þÿ S u o m e n k i e l i s e n t e k o ä l y n k e h i t t ä m i s o h j e l m a e s

Vake Oy
2019

Jauhiainen , T , Lennes , M & Marttila , T (toim) 2019 , Suomenkielisen tekoälyn þÿ k e h i t t ä m i s o h j e l m a e s i s e l v i t y s . V a k e O y , H e l s i n k i . < https://vake.fi/wp-content/uploads/Vake-suomenkielisen-teko%C3%A4lyn-kehitt%C3%A4minen-esiselvitys-2019.pdf >

http://hdl.handle.net/10138/319478

cc_public_domain
publishedVersion

Suomenkielisen tekoälyn kehittämisohjelma – esiselvitys

Toimittaneet: Tommi Jauhiainen (Helsingin yliopisto), Mietta Lennes (Helsingin yliopisto) ja Terhi Marttila (Vake)

Ohjausryhmä: Pia Erkinheimo (Vake), Outi Keski-Äijö (Business Finland), Mikko Kurimo (FCAI/Aalto yliopisto), Krister Lindén (FIN-CLARIN/Helsingin yliopisto), Aki Parviainen (Business Finland), Tuomas Teuri (Vake), Alexander Törnroth (Teknologiateollisuus ry)

1. Esipuhe

”Suomenkielisen tekoälyn kehittämisohjelma"-esiselvitys on Valtion kehitysyhtiö Vaken tilaama esiselvitys, jonka tarkoituksena on tukea suomen kielen laajaa käytettävyyttä erilaisissa tekoälysovelluksissa ja tähän liittyvää palveluiden ja teknologian kehitystä sekä tukea Vaken ynnä muiden organisaatioiden toimintaa tämän mahdollistamiseksi. Esiselvityksen päätarkoituksena on muodostaa toimenpide-ehdotuksia, niin kutsuttuja operaatioita, joiden avulla tekoälyn käyttäminen suomen kielellä on tulevaisuudessa entistä paremmin mahdollista. Toimenpide-ehdotuksista tärkeäksi nousi kieliresurssien kehittäminen uuden toimijan perustamisen tai olemassa olevan toimijan tukemisen kautta. Erityisenä painopisteenä on mahdollistaa avointen yrityskäyttöön soveltuvien suomenkielisten kieliresurssien (esimerkiksi kieliaineistojen ja ohjelmakirjastojen) kehittäminen. Esiselvitys aloitettiin toukokuussa 2019, suoritettiin kesän aikana ja se valmistui syyskuussa 2019.

Toimenpide-ehdotukset kuvataan alustavina projektiaihioina tämän dokumentin viimeisessä osassa. Kaikkien suunniteltujen operaatioiden tavoitteena on luoda vapaasti (avoimen lähdekoodin lisenssillä tms.) saatavissa tai käytettävissä olevia kieliresursseja suomen kielelle. Keskeisimpinä kieliresursseina pidetään kieliaineistoja eli kielikorpuksia, jotka sisältävät ihmisten tuottamaa kieltä puheena ja/tai tekstinä. Tärkeää on löytää tai perustaa toimija, joka ylläpitää, kehittää ja tarjoaa näitä kieliresursseja.

Mahdollisimman laaja-alaisen yhteiskunnallisen vaikuttavuuden varmistamiseksi projekteissa luodut kieliresurssit olisivat kaikkien kiinnostuneiden käytettävissä – eivät ainoastaan niiden, jotka ovat osallistuneet tämän esiselvityksen koostamiseen tai sen pohjalta aloitettujen projektien määrittelemiseen ja toteuttamiseen. Suunniteltujen kieliresurssien toivotaan merkittävästi edistävän mahdollisuuksia suomenkielisen tekoälyn käyttöönottoon erilaisissa sovelluksissa ja palveluissa niin yksityisellä kuin julkisellakin sektorilla.

Esiselvityksen kirjallisen version tarkoituksena ei ole sisältää kattavasti tietoa kaikista kieliresursseihin liittyvistä seikoista vaan pikemminkin nostaa esiin tärkeitä esimerkkejä ja näkökohtia. Joihinkin tarpeisiin ja ongelmakohtiin perehdytään hieman tarkemmin kuin toisiin, mikäli niihin on selvityksen aikana ollut saatavilla asiantuntemusta.

Esiselvityksen ennalta määritelty takaraja tuli käytännössä vastaan hyvin nopeasti. Liitteenä olevasta listasta löytyvät esiselvityksen koostamiseen ja sisällön tuottamiseen osallistuneet tahot ja henkilöt. Listan pituus ja esiselvityksen tekemiseen eri tahojen

puolesta osoitettu tarmo ja mielenkiinto ovat vain kasvaneet aivan selvityksen viime metreille saakka. Monin tavoin tuntuu siltä, että nyt on päästy vasta hyvin vauhtiin ja on sääli lopettaa tämä työ ikään kuin kesken. Ohjausryhmässä olemme kuitenkin yksimielisiä siitä, että esiselvityksessä on nyt kerätty riittävästi tietoa tarkempien operaatioiden, kuten alustatalouden mallia noudattavan toimijan edellytysten mahdollistamisen sekä sitä tukevien muiden projektien aloittamiseksi ja *on aika siirtyä kehittämisohjelman seuraavaan vaiheeseen.*

Kunnia ja kiitokset tästä esiselvityksestä kuuluvat kaikille liitteessä mainituille henkilöille. Toivottavasti ainakin osa näistä erittäin kovan luokan asiantuntijoista löytää voimia osallistua myös tämän esiselvityksen pohjalta aloitettujen projektien tarkempaan määrittelyyn ja toteutukseen.

2. Johdanto

Tässä esiselvityksessä tekoälyllä tarkoitetaan laajasti kaikkia sellaisia kehittyneitä tietojärjestelmiä, jotka kykenevät tekemään monimutkaisia päätöksiä ennalta määräämättömien ulkopuolisten havaintojen perusteella.[1] Tekoäly sisältää niin syväoppimiseen[2] soveltuvien neuroverkkojen kuin myös perinteisempien koneoppimiseen[3] tarkoitettujen algoritmien varaan perustuvat tietojärjestelmät.

Yksi tärkeimmistä poliittisista perusteista tekoälyn kehittämiselle on työn tuottavuuden kasvattaminen. Koneistumisen ja teollistumisen ansiosta esimerkiksi maanviljely tehostui niin, että tällä hetkellä alle 5 % väestöstä hoitaa koko kotimaisen ruokatuotannon. Samaan aikaan voidaan tuoda tavaroita yhä kauempaa, minkä mahdollistamiseksi on ollut tarpeen luoda uusia työpaikkoja ja saada tietoa ulkomaiden tilanteesta. Televiestinnän kehityksen ansiosta viestintä on tehostunut niin paljon, että voimme välittömästi tietää, mitä maapallon toisella puolella tapahtuu. Lähes jokaisella on oma älypuhelin, jolla hän voi osallistua globaaliin kommunikaatioon. Tämä kehitys on tehostanut viestinnän ja markkinoiden toimintaa, mikä puolestaan on kiihdyttänyt globalisaatiota. Samalla tietoverkkojen ja kuljetusalan palvelutehtävät ovat lisääntyneet. Tekoäly ja automaatio tulevat seuraavaksi tehostamaan palveluammatteja, joissa tekoälyä voidaan hyödyntää tiedonkeruussa ja tiedonjalostuksessa.

Tekoälyn käyttöön on myös yhteiskunnallisia ja sosiaalisia perusteita. Tekoälyn avulla on mahdollista tuottaa ja räätälöidä entistä yksilöllisempiä palveluita, joiden avulla voidaan paremmin huomioida myös erityistä tukea vaativia ihmisiä, esimerkiksi vanhuksia, näkö- ja kuulovammaisia, sekä liikuntarajoitteisia henkilöitä. Suurten joukkojen itsepalvelu ja erityisryhmien palvelusovellukset kuitenkin edellyttävät, että käyttöliittymät toimivat luotettavasti tekstin ja puheen avulla, käyttäjän omalla äidinkielellä. Jos ajantasainen puhe- ja kieliteknologinen tuki olisi avoimesti saatavilla ja hyödynnettävissä, tämä tarjoaisi myös yrityksille laajat mahdollisuudet kehittää uusia sovelluksia ja rakentaa palvelukokonaisuuksia.

Suomenkielisellä tekoälyllä tarkoitetaan tekoälyä, joka kykenee käsittelemään suomen kieltä, niin puhuttuna[4] kuin kirjoitettunakin. Suomenkielisen tekoälyn tulee myös kyetä tuottamaan suomenkielistä tekstiä sekä puhetta[5].

[1] https://fi.wikipedia.org/wiki/Tekoäly
[2] https://fi.wikipedia.org/wiki/Syväoppiminen
[3] https://fi.wikipedia.org/wiki/Koneoppiminen
[4] Puheentunnistus (eng. speech recognition): https://fi.wikipedia.org/wiki/Puheentunnistus
[5] Puhesynteesi (eng. speech synthesis): https://fi.wikipedia.org/wiki/Puhesynteesi

Valtakielten, kuten englannin, asema tekoälyn käyttökielenä on vahva ja entisestään vahvistumassa. EU:ssa tiedostetaan englannin valta-asema ja tarve muiden kielten kieliteknologian edistämiseen yhteiskunnan tuella. Käynnissä olevan Digitaalinen Eurooppa -ohjelman kyselyssä[6] todetaankin:

”English currently dominates the digital environment, while other EU languages are under-represented. This gives rise to economic, social and cultural barriers. To ensure that latest technologies are available across all EU languages so as to provide all EU citizens with access to online content and services in their language and all SMEs with latest technologies tuned to their needs, this action will support localisation and deployment of language technologies such as automatic translation, subtitling or text analytics across Europe.”

Tilannetta kuvaa hyvin myös se, että kyseinen kysely on käytettävissä vain englannin kielellä.

Suomenkielisten tekoälysovellusten vähäisyys johtuu osin käytettävissä olevien tekniikoiden ja aineistojen puutteesta, mutta myös yleisesti markkinoiden pienuudesta. Tällä hetkellä näyttäisi siltä, että esimerkiksi ne kielet, joille Google tarjoaa luonnollisen kielen käsittelyyn tarkoitettuja työkaluja, ovat valikoituneet lähinnä markkinoiden laajuuden perusteella.

Tässä esiselvityksessä ja sen suosittelemissa operaatioissa keskitytään erityisesti suomen kieleen ja sen murteisiin[7], mutta mallien luomiseen ja kieliresurssien rakentamiseen rakennettujen välineiden toivotaan soveltuvan myös esimerkiksi suomenruotsiin ja saamelaiskielille.

Euroopan komissio on tänä vuonna julkaissut luotettavaa tekoälyä koskevat eettiset ohjeet.[8] Ohjeistuksen on laatinut tekoälyä käsittelevä korkean tason asiantuntijaryhmä (AI HLEG). Ohjeistuksessa keskitytään tekoälyn peruskomponentteja enemmän määrittelemään tekoälyn käyttöön ja käyttökohteisiin liittyviä eettisiä kysymyksiä, mutta se on osin relevantti myös tekoälyn kielikomponentteja määriteltäessä ja rakennettaessa.

[6] https://ec.europa.eu/digital-single-market/en/news/have-your-say-future-investment-europes-digital-economy
[7] Suomen kieli jaetaan yleisesti kahdeksaan eri murrealueeseen: https://www.kotus.fi/kielitieto/murteet/suomen_murteet
[8] https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60426

Tämän dokumentin kolmannessa luvussa käydään läpi tunnistamiamme vastaavanlaisia kieliresurssien keräämiseen keskittyviä hankkeita muille kielille. Neljännessä luvussa esitetään esimerkinomaisesti muutamia ylemmän tason käyttötapauksia suomenkieliselle tekoälylle ja pohditaan tekoälyn kehittymisen mahdollisia yhteiskunnallisia vaikutuksia. Viidennessä luvussa esitellään joitakin olemassa olevia yrityskäyttöön soveltuvia kieliresursseja, jotka ovat osaltaan vaikuttaneet suunniteltujen operaatioiden sisältöön. Esiselvityksen kuudennessa luvussa esitellään suosituksemme toimenpiteiksi alustavien hankesuunnitelmien muodossa.

3. Vastaavat hankkeet muille kielille

Vastaavanlaisia hankkeita, joissa keskiössä olisi tietyn kielen tarvitsemien kieliresurssien tuottaminen yrityskäyttöön, ei esiselvityksen aikana löytynyt kuin yksi. ”AI INNOVATION of Sweden” toimii Ruotsissa kansallisena tekoälykeskuksena, jonka yhtenä tarkoituksena on nimenomaan yritysten tekoälykehityksen edistäminen tuottamalla tarvittavia resursseja, tietoa ja osaamista. Toukokuun lopulla keskus tiedotti erillisestä projektista, jonka nimenomaisena tarkoituksena on kerätä ja luoda kieliresursseja ruotsin kielelle.[9] Projektin päämäärät vaikuttavat hyvin samankaltaisilta kuin tämänkin esiselvityksen kohteena olevan hankkeen.

Akateemisella puolella vastaavia hankkeita muille kielille löytyi useita. Vuonna 1998 osaksi EU:n viidettä tutkimuksen puiteohjelmaa[10] suunniteltiin ns. BLARK-mallia kielten kieliteknologisten resurssien määrittelyyn. BLARK on lyhenne nimestä ”Basic LAnguage Resource Kit”, eli ”kieliresurssien peruspakkaus”. BLARK toimii eräänlaisena viitekehysmallina, jonka avulla voidaan esittää jonkin kielen kieliteknologisten resurssien kypsyystaso. Steven Krauwer hahmotteli BLARKin alkuvaiheessa seuraavasti:[11]

A. yleinen tekstikorpus, joka vähintään tarvitaan, jotta tiettyä kieltä voidaan ylipäänsä tutkia: vaikkapa 10 miljoonan sanan annotoitu sanomalehtikorpus
B. puhekorpus vastaavasti
C. perustyökalujen kokoelma korpuksen käsittelyä ja analysointia varten
D. kilpailukykyisen luonnollisen kielen tai puheteknologiateollisuuden kehittämistä varten minimissään tarvittavat tiedot ja taidot / tietämys

[9] https://www.ai.se/en/news/new-nlp-project-improve-linguistic-understanding-swedish-ai-applications
[10] https://cordis.europa.eu/programme/rcn/624/en
[11] http://www.speech.kth.se/prod/blark/elsnet%26elra.pdf

Vuoden 2003 artikkelissaan[12] Krauwer antaa hieman pidemmälle viedyn, mutta vain esimerkinomaisen listan:

- kirjoitetun kielen korpuksia
- puhutun kielen korpuksia
- yksi- ja kaksikielisiä sanakirjoja
- terminologioita
- kielioppeja
- ohjelmistokomponentteja (esim. taggereita[13], morfologisia analysaattoreita, jäsentimiä, kielentunnistimia, puhesynteesi)
- standardeja ja välineitä annotaatiota varten
- välineitä korpusten tutkimiseen ja käyttämiseen
- kaksikielisiä korpuksia
- jne.

Vuonna 2002 Binnenpoorte ja kumppanit esittelivät BLARK-mallin hollannin kielelle.[14] Hollannin kielellä on puhujia yli 20 miljoonaa,[15] mikä on vähän verrattuna paljon puhuttuihin kieliin (kiina, hindi, englanti, espanja, arabia, portugali), mutta kuitenkin selvästi enemmän kuin suomen puhujia. Heidän lopullinen listauksensa hollannilta puuttuvista resursseista sisälsi seuraavat kohdat:

A. Tekstiteknologia:
    a. hollannin annotoitu tekstikorpus: puupankki[16] jossa syntaktinen ja morfologinen annotaatio
    b. syntaktinen analyysi: robusti tekstin virkkeiden jäsennys
    c. robusti tekstin esikäsittely: saneistus[17] ja nimettyjen entiteettien havaitseminen
    d. semanttiset annotaatiot edellä mainittuun puupankkiin
    e. verrannolliset korpukset[18]
    f. suorituskykytestit evaluaatiolle
B. Puheteknologia:
    a. automaattinen puheentunnistus (sisältäen toisen kielen puhujien puheen tunnistamisen, robustin puheentunnistuksen ja prosodian[19] tunnistamisen moduulit)
    b. puhekorpuksia tiettyihin loppukäyttötapauksiin

[12] http://elsnet.let.uu.nl/dox/krauwer-specom2003.pdf
[13] https://en.wikipedia.org/wiki/Part-of-speech_tagging
[14] http://www.lrec-conf.org/proceedings/lrec2002/pdf/252.pdf
[15] https://fi.wikipedia.org/wiki/Hollannin_kieli
[16] https://en.wikipedia.org/wiki/Treebank
[17] http://tieteentermipankki.fi/wiki/Nimitys:saneistaa
[18] http://tieteentermipankki.fi/wiki/Käännöstiede:verrannollinen_korpus
[19] https://fi.wikipedia.org/wiki/Prosodia

c. multimediapuhekorpus (puhekorpus, joka sisältää myös informaatiota muussa muodossa kuin puheena)
d. työkalut puheen (puoli)automaattiseen transkriptioon
e. puhesynteesi
f. suorituskykytestit evaluaatiolle

Selvitystä tehdessään he havaitsivat, että sen aikainen hollannin kielen HLT-infrastruktuuri[20] oli hajallaan, epätäydellinen eikä riittävän saavutettavissa. Usein saatavissa olevat kieliresurssit olivat huonosti dokumentoituja. BLARKin komponenttien tulisi myös olla saatavilla edullisesti tai ilmaiseksi. Varsinaisina jatkotoimenpiteinä he ehdottivat, että:

A. jo olemassa olevat resurssit pitäisi koota, dokumentoida ja ylläpitää jonkinlaisessa HLT-organisaatiossa.
B. BLARKin puutelistalla olevat resurssit pitäisi toteuttaa rohkaisemalla rahoittajatahoja tukemaan puuttuvien resurssien kehittämistä.
C. BLARK-komponenttien pitäisi olla tutkijoiden ja HLT-yritysten saatavilla avoimilla lisensseillä.
D. suorituskykytestit, testikorpukset ja objektiivisen vertailun metodologia, evaluaatio ja validaatio BLARKin eri komponenteille pitäisi kehittää.

Lopuksi he vielä totesivat, että hollannin koulutetuista HLT-osaajista oli puute ja että rahoitusta pitäisi suunnata riittävästi myös perustutkimukseen.

Vuonna 2006 Maegaard ja kumppanit[21] julkaisivat ensimmäisen version arabian kielen BLARKista. Siinä he päätyivät hyväksymään mukaan myös maksullisia resursseja. He jakoivat resurssit neljään eri hintaluokkaan:

1. yli 10 000 €
2. 1000 – 10 000 €
3. 100€ – 1000 €
4. alle 100 € tai ilmainen

Yli 40 000 € maksavia resursseja hekään eivät hyväksyneet mukaan. BLARKin pohjalta he päättivät tuottaa kolme kieliresurssia:

1. noin 500 000 sanan kirjoitetun kielen korpuksen
2. kaksi viiden tunnin puhekorpusta puhesynteesille
3. noin 40 tunnin puhekorpuksen uutisista yleisarabiaksi[22]

[20] Human Language Technologies
[21] http://lrec-conf.org/proceedings/lrec2006/pdf/521_pdf.pdf
[22] Modern Standard Arabia (MSA)

Vuonna 2017 Hossein Hassani julkaisi BLARKin kurdin kielelle.[23] Hassani huomauttaa, että aikaisemmat BLARKit on suunniteltu vain kielille, joissa ei ole paljon murrevaihteluita. Tämä ei pidä paikkaansa, sillä todellisuudessa kaikissa kielissä lienee paljon murteenkaltaista paikallista ja/tai eri yhteisöjen välistä vaihtelua. Esimerkiksi arabian kielessä on paljon murrevaihtelua, mutta käytännössä vuoden 2006 BLARK kuitenkin keskittyi lähinnä yleisarabiaan. Kysymys on lähinnä siitä, missä määrin murteita halutaan ottaa mukaan ja missä määrin kielellä on yksi vallitseva virallinen muoto. Kurdin kieli ei ole minkään valtion ensimmäinen virallinen kieli, joten sen kehittymistä ja käyttämistä säätelemään ei ole muodostunut erillistä virallista tahoa. Lähinnä sellaisena toimii e-NGO[24] Kurdish Academy of Language (KAL).[25] Myös kurdin kielessä on useita murteita, ja sitä kirjoitetaan käyttäen neljää eri kirjoitusjärjestelmää, eikä käytössä ole yhtä standardoitua ortografiaa.

Vuonna 2014 Erhard Hinrichs kirjoitti yhdessä Steven Krauwerin kanssa artikkelin,[26] jossa he kuvaavat vuonna 2012 perustettua CLARIN ERICiä.[27] CLARIN pyrkii tarjoamaan tutkijoille helpon ja pysyvän tavan saada käyttöönsä digitaalisessa muodossa olevaa kielimateriaalia sekä sen tutkimiseen, analysointiin ja muuhun hyödyntämiseen tarkoitettuja kehittyneitä välineitä. Työ CLARINin luomiseksi aloitettiin valmistelevan vaiheen projektilla jo vuonna 2008.[28] Suomi oli aktiivisesti mukana, ja FIN-CLARIN[29] perustettiinkin hyvin varhaisessa vaiheessa.

FIN-CLARIN on vuodesta 2009 lähtien kuulunut Suomen Akatemian ylläpitämään suomalaisten tutkimusinfrastruktuurien kansalliseen tiekarttaan.[30] FIN-CLARIN-konsortion muodostavat yhdessä kaikki kielitieteellistä tutkimusta harjoittavat suomalaiset yliopistot[31], Kotus[32] sekä CSC[33]. FIN-CLARIN ylläpitää Kielipankkia,[34] jonne konsortion jäsenet yhdessä tuottavat erilaisia kieliresursseja, niin aineistoja[35] kuin työkalujakin[36].

---

[23] https://link.springer.com/content/pdf/10.1007%2Fs10579-017-9400-0.pdf
[24] ”electronic non-governmentsl organization”
[25] http://kurdishacademy.org/
[26] https://dspace.library.uu.nl/handle/1874/307981
[27] ”Common Language Resources and Technology Infrastructure” ”European Research Infrastructure Consortium”
[28] https://www.clarin.eu/content/about-clarin-preparatory-phase
[29] https://www.kielipankki.fi/organisaatio/fin-clarin/
[30] https://www.aka.fi/globalassets/tiedostot/aka_infra_tiekartta_raportti_fi_030518.pdf
[31] Helsingin yliopisto, Aalto-yliopisto, Itä-Suomen yliopisto, Jyväskylän yliopisto, Oulun yliopisto, Tampereen yliopisto, Turun yliopisto sekä Vaasan yliopisto.
[32] https://www.kotus.fi
[33] https://www.csc.fi
[34] https://www.kielipankki.fi
[35] https://www.kielipankki.fi/aineistot/
[36] https://www.kielipankki.fi/tyokalut/

Alusta alkaen sekä aineistot että työkalut on ajateltu saatettavan nimenomaan tutkimusyhteisön käyttöön mahdollisimman tehokkaalla tavalla. Koska aineistojen ja työkalujen yrityskäyttöön saattamisen tarpeeseen ei ole kiinnitetty huomiota, on useat kieliresurssit lisensoitu nimenomaan NC - non-commercial - rajauksella. Monissa tapauksissa tekijänoikeudelliset tai tietosuojaan liittyvät kysymykset ovat lisäksi aiheuttaneet rajoituksia lisensseihin. FIN-CLARINissa CSC vastaa kokonaisuuden teknisestä ylläpidosta ja Helsingin yliopisto aineistojen ja työkalujen hankinnasta sekä koulutustoiminnasta.

Toinen tällä hetkellä aktiivinen EU:n laajuinen hanke on ELRC[37], joka on keskittynyt erityisesti konekääntämiseen tarvittavien kieliresurssien keräämiseen ja tuottamiseen EU:n virallisille kielille. Kolmas EU:n laajuinen meneillään oleva hanke on ELG[38] (2019 – 2021), joka pyrkii edistämään kieliteknologian käyttömahdollisuuksia ja löydettävyyttä keräämällä kieliresursseja yhteiselle alustalle.

4. Kieliresurssien tarve ja niiden kehittämisen odotetut vaikutukset

Tässä luvussa tuodaan esimerkinomaisesti esille joitakin erilaisia loppukäyttötapauksia, jotka nousivat esille esiselvitykseen osallistuneiden tahojen kanssa käydyissä keskusteluissa.

Yrityksillä on tuotantokäytössä useita suomenkielistä tekoälyä hyödyntäviä järjestelmiä. On kuitenkin useita käyttötapauksia, joiden järkevä tuotteistaminen ei nykyisillä kieliresursseilla ole onnistunut. On hyvin vaikea arvioida, miltä osin tämä johtuu siitä, etteivät kieliresurssien kysyntä ja tarjonta eivät välttämättä aina kohtaa. Esiselvityksen aikana käytyjen keskustelujen perusteella suurin osa esiselvityksen koostamiseen osallistuneista asiantuntijoista kuitenkin arvioi, että avoimesti saatavilla olevat kieliresurssit vauhdittaisivat erilaisten suomenkielisten tekoälysovellusten rakentamista ja käyttöönottoa.

Avoimesti saatavilla olevilla komponenteilla olisi mahdollista esimerkiksi pilotoida tiettyä ylemmän tason palvelua, minkä jälkeen olisi mahdollista ottaa käyttöön kaupallisesti tarjolla oleva, kokonaan eri tekniikkaan perustuva tai samasta avoimesta komponentista edelleen kehitetty tuote.

[37] European Language Resource Coordination: http://www.lr-coordination.eu
[38] European Language Grid: https://www.european-language-grid.eu/about/

4.1. Puhe

Monia suomenkielistä puhetta käyttäviä palveluita ja järjestelmiä on jo tuotantokäytössä. Myös suomen kielellä toimiva automaattinen puheentunnistus (puhe tekstiksi -toiminnallisuus, engl. *speech-to-text, STT*) ja puhesynteesi (tekstistä puheeksi, engl. *text-to-speech*, *TTS*) ovat jo useiden vuosien ajan olleet käyttäjien saatavilla erilaisissa päätelaitteissa ja sovelluksissa (esim. Applen ja Googlen puhekyvykkyydet). Monien palveluiden toteuttaminen tai tehostaminen edellyttäisi kuitenkin nykyistä parempaa ja luotettavampaa suomenkielisen puheen tukea.

4.1.1. Vapaamuotoisten keskusteluiden tallenteiden saattaminen tekstimuotoon

Yksi vaikeimpia suoraan puhetallenteisiin liittyviä asioita on haku puheen sisällöstä. Mikäli puhe muutetaan tekstimuotoon, voidaan siihen käyttää jo olemassa olevia tekstimassojen analyysiin ja louhintaan tarkoitettuja välineitä. Yksinkertaisimmillaan halutaan esimerkiksi tietää, monessako puhelussa on mainittu jokin tutkittava asia. Esimerkkejä monimutkaisemmista käyttötapauksista ovat erilaiset analyysit puhelun kulusta sekä puheluiden tekoälypohjaiset jälkikäsittelyratkaisut.

Toinen käyttökohde on esimerkiksi toimittajien tekemien haastattelujen automaattinen litterointi. Pelkästään se, että puheesta voisi litteraation perusteella nopeasti löytää tietyn kohdan, nopeuttaisi haastattelujen yksityiskohtien tarkistamista selvästi. Tällaisessa tapauksessa tekstin asemointi kappaleiksi tai puheenvuoroiksi ja sen nopea luettavuus ja silmäiltävyys ovat tärkeämpiä kuin tekstin täydellinen vastaavuus puheen kanssa. Puhetallenteiden hakumahdollisuuksien parantamisesta olisi hyötyä myös virkamiestyönä tehtävässä faktantarkistuksessa, kun halutaan esimerkiksi tarkistaa eduskunnan täysistunnoissa tehtyjä päätöksiä istunnoista tehtyjen videotallenteiden pohjalta.

Mahdollisuuksia on paljon, mutta puheentunnistuksen tämänhetkinen laatu estää niistä monien toteuttamisen. Ongelmia esiintyy etenkin monen puhujan välisessä vapaassa keskustelussa ja tilanteissa, joissa esiintyy runsaasti taustahälyä, jolloin virheellisten tunnistustulosten riski on vielä tätä nykyä suuri.

4.1.2. Sanelu ja tekstitys

Automaattista puheentunnistusta tarvitaan ja käytetään paljon perinteiseen tekstin saneluun, vaikkapa viestien alustavaan kirjoittamiseen tilanteissa, joissa halutaan toimia kädet vapaana. Monia käyttötarkoituksia silmällä pitäen tietyn henkilön puheeseen mukautuva sanelukirjoitus toimii jo nykyisin kohtuullisesti myös suomeksi esimerkiksi

mobiililaitteissa; etenkin olosuhteissa, joissa hälyn määrä on vähäinen ja/tai puhuja on lähellä laitteen mikrofonia. Kyseinen toiminnallisuus ei yritysten haastatteluissa noussutkaan erityisesti esille.

TV-ohjelmia, luentoja ym. esityksiä voidaan myös tekstittää automaattisesti. Tekstityksen laadun parantamiseksi käytetään usein ns. sanelutekstitystä (engl. *respeaking*), jossa tietty henkilö kuuntelee ja toistaa tunnistinta varten selkeästi ääneen esimerkiksi tv-ohjelman puheen sisällön. Näin automaattinen tunnistus voidaan sovittaa sanelijan puheelle ja se toimii luotettavammin. Myös sanelutekstityksen haasteena on vielä tällä hetkellä tasapainoilu tunnistuksen nopeuden ja laadun välillä. Tekstityksen käyttökohteesta ja laatuvaatimuksista riippuen sanelijan olisi puhuttava mieluiten selkeää kirjakieltä ja mahdollisesti pyrittävä lennossa tiivistämään puheen sisältöä[39], mikä voi vaatia paljon harjoittelua ja/tai erillistä koulutusta.

Erityisryhmistä esimerkiksi kuulovammaisille olisi paljon apua puheen lähes reaaliaikaisesta tai nopeasta tekstityksestä. Luotettavasti toimiva, tyylilajista riippumaton suomenkielisen puheen tekstitys tarjoaisi myös pohjan esimerkiksi automaattiselle tulkkaukselle, jolla on globalisoituvassa maailmassa lukemattomia käyttökohteita.

4.1.3. Hyvältä kuulostavan suomenkielisen puheen tuottaminen (puhesynteesi)

Puhesynteesi on ihmisen puheen tuottamista keinotekoisesti.[40] Yleisin puhesynteesin muoto on kirjoitetun tekstin muuttaminen puhuttuun muotoon (engl. text-to-speech synthesis). Puhesynteesiä käytetään tällä hetkellä esimerkiksi suurimpien mediayhtiöiden kännykkäsovelluksissa kirjoitettujen uutisten ääneen lukemiseen. Tällä hetkellä yrityksissä käytetään tuotannossa esimerkiksi hollantilaisen ReadSpeakerin[41] SaaS[42]-palvelua tai Microsoftin Azure-pilvipalvelun tekstistä puheeksi -sovellusta[43]. Näiden palveluiden tuottama puhesynteesi on kuitenkin laadultaan vielä varsin kaukana ihmisen tuottamasta puheesta, ja käyttäjien toiveena olisi vielä luonnollisemmalta kuulostava synteettinen puhe. Suomessakin puhesynteesillä on pitkät perinteet, ja esimerkiksi Timehouse Oy on myynyt mikrotietokoneisiin asennettavissa olevaa puhesynteesiä jo 1990-luvulla.[44]

---

[39] Ks. esim. Fröberg, Essi (2018), *Tekstitys ja sanelu. Kuinka tuottaa laadukas kielensisäinen tekstitys suoriin tv-lähetyksiin?* Pro gradu -työ, Helsingin yliopisto. http://urn.fi/URN:NBN:fi:hulib-201806132502
[40] https://fi.wikipedia.org/wiki/Puhesynteesi
[41] https://www.readspeaker.com/about-us/
[42] SaaS: “Software as a Service”
[43] https://www.tekniikkatalous.fi/uutiset/robotti-lukee-aaneen-uutisia-suomenkielinen-puhe-tulee-pilvesta-ja-aanisyntetisaattorille-voidaan-opettaa-kielen-erikoisuuksia/a1d0078d-d023-3529-9f67-3ffd6c17a892
[44] https://www.tivi.fi/uutiset/mikropuhe-syntetisaattorista-uusi-versio/924f75d9-4388-37da-bd1d-b22e563c9f06

4.1.4. Automaattiset avustajat ja muut puhepohjaiset käyttöliittymät

Nyky-yhteiskunnassa toimiminen vaatii käyttäjiltä paljon ”digitaitoja” ja esimerkiksi mobiililaitteiden hallintaa. Jos esimerkiksi käyttäjän näkökyky on heikentynyt tai sorminäppäryys ei riitä laitteen käyttämiseen, hän saattaa jäädä monien palveluiden ulkopuolelle. Usein digitaitovaatimukset voitaisiin kuitenkin ohittaa, mikäli saatavilla olisi käyttäjän äidinkieltä osaava, puheella toimiva käyttöliittymä. Puheella toimivat älykkäät sovellukset voisivat täydentää tai jopa korvata esimerkiksi vanhuksille ja vammaisille tarjottavia henkilökohtaisia palveluita, tai esimerkiksi tarjota mahdollisuuden apua tarvitsevien henkilöiden itsenäiseen ja omatoimiseen asumiseen ja siten parantaa heidän elämänlaatuaan. Toisaalta, jos puhekäyttöliittymä on olemassa mutta toimii huonosti, se herättää epäluottamusta ja asiakas alkaa kenties välttää palvelun käyttöä. Tietyissä tapauksissa, mm. terveydenhuollon palveluissa, käyttöliittymän puutteet voivat myös aiheuttaa turvallisuusriskejä.

Kehittyneitä, käyttäjäkohtaisesti mukautuvia puhekäyttöliittymiä voitaisiin hyödyntää nykyistä paremmin myös erilaisissa kieltenopiskelua tukevissa sovelluksissa. Suullista kielitaitoa ja vuorovaikutustaitoja korostetaan yhteiskunnassa ja työelämässä yhä enemmän ja ne nähdään olennaisena osana niin oman äidinkielen kuin vieraan kielenkin osaamista. Suomeen tulleiden maahanmuuttajien kotoutumisessa etenkin hyvä suullinen suomen kielen taito on suureksi eduksi työllistymisen ja sosiaalisten verkostojen muodostumisen kannalta. Myös esimerkiksi ylioppilastutkintoon sisältyviin kielten kokeisiin on ollut pitkään suunnitteilla kielen suullisen taidon tasoa mittaava osio. Päänvaivaa kuitenkin aiheuttavat edelleen yhtäältä digitaalisten koetehtävien tekniset rajoitukset ja toisaalta suullisen kielitaidon arviointikriteereiden yhteismitattomuus. Kenties molemmat ongelmat kannattaisikin yrittää ratkaista yhtä aikaa.

Suullisen kielitaidon arviointiin ei riitä automaattinen puheesta tekstiksi -sovellus, vaikka sellainen sinänsä toimisikin luotettavasti. Kielenoppijan ääntämistä tai hänen toimintaansa vuorovaikutustilanteessa ei ole mahdollista mitata kirjoitetun tekstin eikä hänen puheestaan tehdyn litteroinnin tai edes tarkimmankaan foneettisen transkription perusteella. Transkriptiojärjestelmästä riippumatta puheen siirtäminen tekstimuotoon on aina subjektiivinen prosessi: lopputulos riippuu tekstin tai transkription käyttötarkoituksesta, transkription tekijän omista havainnoista sekä niistä piirteistä, joita hän tietoisesti tai tiedostamattaan haluaa nostaa esiin. Kirjoitetusta tekstistä jää pois kirjaimellisesti lukemattomia luonnollisen vuorovaikutuksen ja puheen merkityssisällön kannalta olennaisia asioita: esimerkiksi äänenpainot ja -sävy, puheen rytmi ja tauotus suhteessa keskustelukumppaneihin, sävelkulku, eleet, ilmeet, katseet ja muu vuorovaikutustilanteeseen ja kontekstiin liittyvä toiminta. Kielenopiskeluun tarkoitetuissa sovelluksissa pelkkä teksti ei siis riitä kuvaamaan sitä, mitä oppija osaa tai mitä hänen pitäisi osata aidossa puhetilanteessa.

Puhutun kielen harjoittelu ja automaattinen testaaminen edellyttävät analytiikkaa suoraan käyttäjän tuottamasta puhesignaalista, joskus myös videokuvasta. Koska jokainen puhuja on fyysisiltä ominaisuuksiltaan ja siis myös puheääneltään yksilöllinen, puhetuotoksen arviointikriteereiden pitäisi joustavasti mukautua kunkin käyttäjän mukaan, mutta niiden pitää olla myös objektiivisesti vertailukelpoisia muiden arvioitavaan ryhmään kuuluvien kanssa.

Esimerkiksi kielenopetussovelluksessa tapahtuvan ääntämisen arvioinnin lisäksi puhekäyttöliittymissä on monia muitakin käyttötapauksia, joissa analysoitavaa puhetta ei ole edes tarpeen esittää tekstimuodossa vaan puheesta voidaan tehdä päätelmiä suoraan. Tällaisia toiminnallisuuksia voisivat olla esimerkiksi puhujan automaattinen tunnistaminen (esimerkiksi puhujan ääni ylimääräisenä verifikaatiomenetelmänä) tai vaikkapa käyttäjän iän, vireystilan tai terveydentilan automaattinen analyysi. Jälkimmäiset voisivat olla sovellusten mukauttamisen ja erilaisten käyttäjälle tarjottavien palveluiden kannalta hyödyllisiä, vaikkei tunnistustarkkuus olisikaan sataprosenttinen.

Tietoturvallisten, yksityisyyden suojaa kunnioittavien puhekäyttöliittymien rakentaminen saattaa joissakin tapauksissa edellyttää, että sovelluksen komponentteja on mahdollista hyödyntää ilman tietojen siirtoa tai luovutusta palvelusta toiseen, kolmansille osapuolille tai toiseen valtioon. Nämä tekijät puoltavat sitä, että käyttöliittymän komponenttien tulisi olla avoimesti saatavilla, tarkasteltavissa ja tarpeen mukaan paketoitavissa ja yhdisteltävissä uudelleen.

4.2. Teksti

Monella yrityksellä on käytössään kieliteknologisia komponentteja suomenkielisen tekstin käsittelyä varten. Osaan näistä ollaan hyvin tyytyväisiä, mutta selvä tarve on olemassa myös tekstin automaattisen käsittelyn parantamiselle. Moni näistä olemassa olevista komponenteista toimii hyvin virkakielelle tai ns. yleiskielelle. Suomalaiset kuitenkin käyttävät internetissä runsaasti erilaisia kirjoitustyylejä ja kieltä, jossa esiintyy monenlaisia puhekielisiä ja murteellisia piirteitä. Näiden kohdalla monet yleisesti käytössä olevat kielikomponentit tuottavat heikkolaatuisia analyyseja. Myös käyttäjän puutteellinen suomen kielen taito vaikuttaa hänen mahdollisuuksiinsa käyttää kielityökaluja, jotka on suunniteltu yleiskieliselle tekstille.

4.2.1. Tekstin monimuotoinen sisältöanalyysi

Tarvitaan helposti saatavia erilaisia tekstin sisällön analysointiin tarkoitettuja työkaluja, esimerkiksi:

- NER (named entity recognition)[45] eli nimien ja niiden kaltaisten rakenteiden havaitseminen tekstistä
- sentimenttianalyysi eli tekstin sävyn päätteleminen negatiivinen-positiivinen-akselilla

Tällaisen tekstianalyysipalvelun asiakkaina toimivat erilaiset tutkimustalot, asiakasymmärrysyritykset sekä viestintä- ja markkinointitoimistot. Viestinnässä ja markkinoinnissa on tärkeää esimerkiksi se, etteivät lentojen mainokset asetu samaan yhteyteen lentämiseen negatiivisesti suhtautuvan uutisen kanssa (esimerkiksi lento-onnettomuuden).

4.2.2. Tekstin luokittelu

Tekstin sisältöanalyyttisia välineitä voidaan käyttää yhdessä yleisimpien koneoppimismenetelmien kanssa luokittelemaan tekstiä eri kategorioihin ja esimerkiksi klusteroimaan samankaltaisia tekstejä. Tämä on tärkeää esimerkiksi uutisia kirjoitettaessa, jolloin ohjelmallisesti voidaan ehdottaa kirjoittajalle referenssiksi aikaisempia samankaltaisia artikkeleita tai vaikkapa samaan aihepiiriin liittyviä asiakasviestejä.

Tekstin luokittelua voidaan myös käyttää asiakaspalveluviestien tärkeyden määrittelemiseen. Hyvin toimivan automaattisen luokittelun avulla on mahdollista automaattisesti päättää, mitkä viestit vaativat toimenpiteitä muita nopeammin, mikä puolestaan lyhentää vakavien ongelmatilanteiden ratkaisuun tarvittavaa aikaa. Tästä on hyötyä sekä yrityksen maineelle että asiakkaille. Tekoälyn avulla on myös mahdollista alustavasti luokitella avoimia palautteita enemmän ja vähemmän hyödyllisiin, jolloin rakentavammat palautteet saavat enemmän huomiota palautteenhallintaan käytetyssä järjestelmässä.

4.2.3. Tekstin koneellinen tuottaminen

Yksi esimerkki tekstin koneellisesta tuottamisesta on uutisten tuottaminen automaattisesti. Automaattisesti tuotettuja uutisia voidaan räätälöidä hyvin tarkoille kohderyhmille: vaikkapa paikallisten urheiluseurojen tuloksista voidaan tuottaa uutisia nimenomaan paikallisille lukijoille. Esimerkiksi Vasabladet käyttää ruotsalaisen teknologiayrityksen United Robotsin tekniikkaa tähän tarkoitukseen.[46] Vastaavanlaiselle tekniikalle myös suomeksi olisi tarve. Tälläkin hetkellä tämänkaltaisia automaattisen tekstin tuottamisen järjestelmiä on tuotantokäytössä, mutta ne perustuvat lähinnä valmiiden lomakepohjien automattiseen täyttämiseen.

[45] https://en.wikipedia.org/wiki/Named-entity_recognition
[46] https://www.vasabladet.fi/Artikel/Visa/122192

4.2.4. Konekääntäminen

Konekääntämisen laadun parantamista ei koettu tärkeäksi suomenkielisestä tekoälystä kiinnostuneita yrityksiä haastatellessa. Osin konekäännöksen huono taso toimii esimerkiksi eräänlaisena markkinasuojana ulkomaanuutisille: mikäli konekäännös toimisi virheettömästi, voisi uutisen lukea suoraan ulkomaisesta palvelusta.

Toisaalta konekäännöstä käytetään myös osana laajempia kieliteknologisia järjestelmiä tuottamaan esimerkiksi suomenkielistä lisäinformaatiota.[47]

4.2.5. Automaattinen tekstintunnistus

Automaattinen tekstintunnistus (OCR[48]) on käytössä useilla tahoilla. Koneellisesti tuotetulle juoksevalle tekstille se toimii nykyään oikein hyvin, mutta esimerkiksi käsin kirjoitetut tekstit, joista on lähetetty kännykkäkuva jonkin www-lomakkeen liitteeksi, tuottavat edelleen ongelmia. Myöskään numeroiden asettelu järkevästi tekstin lomaan ei välttämättä toimi tekstintunnistuksessa täydellisesti. Tämä ongelma lienee tosin käytetystä kielestä riippumaton.

4.3. Muut

Puhe- ja tekstikomponenttien lisäksi on myös muita tarpeita, joita voidaan pitää kieliresursseina. Yksi tällainen olisi esimerkiksi vakaasti ylläpidetty tietopankki, josta olisi mahdollista löytää suomenkieliseen tekoälyyn liittyvää informaatiota, aineistoja ja työkaluja. Kynnys yleisten koneoppimismenetelmien soveltamiseen suomen kielen käsittelyssä on hyvin matala. Tällöin edetään helposti tuntematta lainkaan kieleen liittyviä erityispiirteitä tai sen parissa jo tehtyä tutkimusta. Esimerkiksi nykyiset Kielipankin kautta saatavat aineistot[49] ja ohjelmistot[50], joita yritykset voisivat käyttää kaupallisissa tuotteissaan, hukkuvat helposti yksinomaan tutkimustarkoituksiin rajattujen resurssien joukkoon eivätkä silloin päädy yrityskäyttöön.

Kieliresursseihin, niiden keräämiseen, jakeluun ja käyttämiseen liittyvät laki- ja sopimusasiat nousivat myös korkealle prioriteetille yrityshaastatteluissa. Erityisesti puheaineistojen osalta tietosuoja-asiat ja GDPR-lainsäädäntö monimutkaistavat näiden aineistojen käyttämistä eikä selviä pelisääntöjä ole vielä alalle kehittynyt. Ongelma on yhteinen sekä pienille että isoille yrityksille. Isoilla yrityksillä on yleensä käytettävissään

[47] https://www.tekniikkatalous.fi/uutiset/ssab-ottaa-tekoalyn-tyosuojelun-avuksi-tavoite-olla-maailman-turvallisin-terasyhtio/ca5fb9a5-2bd2-33f7-95c9-8eb9accae100
[48] “Optical Character Recognition”: https://fi.wikipedia.org/wiki/Tekstintunnistus
[49] https://www.kielipankki.fi/aineistot/
[50] https://www.kielipankki.fi/tyokalut/

enemmän sopimus- ja lakiasiantuntemusta, mutta isojen yritysten kohdalla toisaalta myös riskit ovat suurempia.[51]

4.4. Suomenkielisen tekoälyn kehittämisen yhteiskunnallinen vaikuttavuus

Yksi suurimmista yhteiskuntaan yleisesti kohdistuvista tekoälyn kehittymisen vaikutuksista on erilaisten palveluiden automatisointi ja mahdollisuus räätälöidä niitä ihmisten yksilöllisiin tarpeisiin. Kehittyneen automaation lisäksi suomenkielistä tekoälyä hyödyntävistä järjestelmistä saattaa syntyä myös täysin uusia työkaluja, jotka voivat lisätä kansalaisten hyvinvointia sekä työntekijöiden tuottavuutta ja tehoa, perusoikeuksista samalla huolehtien.

Erityisesti vaikutukset tulevat näkymään terveys- ja sosiaalipalveluissa, joissa vuonna 2018 työskenteli lähes puoli miljoonaa henkilöä. Alan henkilöstömäärä on kasvanut noin 28 % vuodesta 2000 vaikka väestön kokonaismäärä on samana aikana kasvanut vain 7 %.[52] Jotta myös vanhukset ja terveydentilaltaan heikentyneet ihmiset kykenisivät kommunikoimaan tekoälyä hyödyntävien kehittyneiden laitteiden kanssa, tulisi niiden kyetä ymmärtämään ihmisten puhetta hyvin tarkasti ja luotettavasti sekä tuottaa helposti ymmärrettävää puhetta. Luontevimmin ja luotettavimmin viestintä onnistuu ihmisten omalla äidinkielellä.

Kansalaisten kielelliset oikeudet määritellään Suomen laissa. Kansalaisten perusoikeuksiin kuuluu muun muassa tasavertaisuus. Oikeusministeriön kanta on, että kielellisten oikeuksien toteutuminen on edellytys muiden oikeuksien toteutumiselle.[53] Kielellisestä saavutettavuudesta Oikeusministeriön vuoden 2018 syyskuussa julkaisemassa selvityksessä[54] kerrotaan seuraavasti:"…kielellistä oikeutta laajempi käsite onkin kielellinen saavutettavuus, joka tarkoittaa sitä, että asiakas tai potilas on tietoinen palveluista, saa palvelua, tulee ymmärretyksi, ymmärtää myös hoitoa koskevat ohjeet ja kykenee siten ottamaan itse vastuuta hoidostaan. Tämä edellyttää mm. tiedotuksen, ohjauksen ja neuvonnan monikanavaisuutta, tulkkipalveluiden saatavuutta ja sähköisten palveluiden saavutettavuutta."

Myös Suomen ja Viron hallitusten yhteisessä julkilausumassa[55] keväältä 2018 korostetaan pienten kielten roolia ja mainitaan seuraavasti: "Tekoälyyn perustuvissa

[51] http://www.enforcementtracker.com
[52] https://www.tilastokeskus.fi/tup/suoluk/suoluk_tyoelama.html
[53] https://oikeusministerio.fi/artikkeli/-/asset_publisher/mita-kielelliset-oikeudet-ovat-
[54] http://urn.fi/URN:ISBN:978-952-259-710-6
[55] https://valtioneuvosto.fi/artikkeli/-/asset_publisher/10616/suomen-ja-viron-hallitukset-sopivat-yhteistyon-syventamisesta

julkisissa palveluissa on varmistettava pienten kielten, kuten viron, suomen ja ruotsin asema."

Palveluiden räätälöinnin ja yleisen saavutettavuuden parantumiselle on mahdotonta määritellä rahallista arvoa, mutta vaikkapa kehittyneen tekoälyn ohjaamien henkilökohtaisten apuvälineiden (esimerkiksi erilaisten siivous- ja ruoanlaittorobottien) voitaisiin olettaa tulevaisuudessa vastaavan yhden henkilön työpanosta jokaisen kotihoidossa olevan vanhuksen kohdalla. Säännöllisessä kotihoidossa oli vuoden 2018 marraskuussa 73 563 henkilöä.[56] Mikäli yhden henkilön työpanosta vastaava lisähyöty saavutettaisiin tasaisesti vuoteen 2030 mennessä, olisi tuotettu lisäarvo pelkästään ensimmäisen kymmenen vuoden aikana noin 1,1 miljardia euroa ja sitä seuraavien 20 vuoden aikana 4,4 miljardia euroa, olettaen että kotihoidon tarpeessa olevien ihmisten määrä pysyisi samana. Ratkaisut syntyvät todennäköisesti jollain aikataululla myös ilman tässä selvityksessä ehdotettuja toimenpiteitä, mutta mikäli ehdotetut toimenpiteet nopeuttavat teknologioiden käyttöönottoa esimerkiksi viidellä vuodella, on laskettu lisäarvo edelleenkin yli miljardi euroa.

Vastaavanlaiset tekoälyyn perustuvat henkilökohtaiset tukiälyt voisivat tuottaa merkittävää hyötyä muillekin ihmisille kaikissa heidän elämänvaiheissaan.[57] Tällöin tuotetun lisäarvon voidaan katsoa olevan moninkertainen edellä laskettuun nähden. Laadukas ihmisen äidinkielellä toimiva palvelu tukee myös hänen henkistä hyvinvointiaan, työtään ja arkeaan.

Terveys- ja sosiaalipalveluita tuottamaan kehitetyt tekoälyn ohjauksessa olevat laitteet ja niihin tarvittavat teknologiat ovat yhteisiä koko maailmalle. Tekoälyn ymmärtämä ja tuottama kieli jää suomen kielen tapauksessa lähinnä suomalaisen yhteiskunnan ja yritysten vastuulle.[58] Kansainvälisesti kehitettyjen tekoälyratkaisujen hyödynnettävyys Suomessa ja suomen kielellä vaatii kehittyneitä suomenkielisiä kieliresursseja.

[56] https://thl.fi/fi/tilastot-ja-data/tilastot-aiheittain/ikaantyneet/kotihoidon-asiakkaat
[57] https://svenska.yle.fi/artikel/2019/08/25/bade-naringslivet-och-facket-eniga-foretag-borde-satsa-mer-pa-artificiell
[58] http://www.meta-net.eu/whitepapers/e-book/finnish.pdf

5. Olemassa olevat avoimet kieliresurssit

Tähän lukuun on esimerkinomaisesti kerätty jo olemassa olevia kaupallisesti hyödynnettäviä avoimia kieliresursseja.

Suomen kielelle löytyy hyvin erilaisia resursseja ja niiden käytettävyys yrityskäyttöön määräytyy lähinnä yrityksen oman kyvykkyyden kautta. Hyvin harva kieliteknologinen kieliresurssi suomen kielelle on käytettävissä ilman erillistä kieliteknologista osaamista. Monet resurssit ovat nimellisesti olemassa, mutta ne eivät kuitenkaan ole käyttökelpoisia tai vain tarpeeksi hyviä tuotantokäyttöön. Osa käytössä olevista avoimista lisensseistä sopii huonosti yhteen yrityskäytön kanssa. Helposti yrityskäyttöön sopivia ovat MIT-[59] ja Apache[60] 2 -lisensseillä jaetut resurssit. Huonosti yrityskäyttöön sopivasta avoimesta lisenssistä esimerkkinä on GNU GPL[61] -lisenssi. GNU Lesser General Public License (LGPL) -lisenssillä[62] jaetut resurssit ovat joskus käyttökelpoisia, mutta yrityksen järjestelmäkokonaisuus saattaa vaatia niiden osalta erityishuomiota toimintaa kehitettäessä: pitää muistaa mitä järjestelmällä saa tehdä ja mitä ei.

5.1. Tekstikorpukset

Yrityskäyttöön suoraan käytettävissä olevia tekstikorpuksia löytyy jo tällä hetkellä Kielipankista, joista tässä alla muutamia esimerkkejä:[63]

- Korkeimman oikeuden ja Korkeimman hallinto-oikeuden päätöksiä vuosilta 1980 – 2018 suomeksi, latausversio[64]
    - Korkeimman oikeuden (KKO) päätöksiä vuosilta 1980 – 2018 suomeksi ja Korkeimman hallinto-oikeuden (KHO) päätöksiä vuosilta 1987 – 2018 suomeksi.
    - KKO:n päätöksiä on 5651 ja KHO:n päätöksiä 7633. Suurimmassa osassa päätöksistä oikeudenkäynnin kieli on ollut suomi. Tällöin dokumentti sisältää koko päätöstekstin. Jos oikeudenkäynnin kieli on ollut ruotsi, dokumentti sisältää pelkän tiivistelmän suomeksi.
- Eduskunnan alkuperäissäädöksiä vuosilta 1734 – 2018, latausversio[65]
    - Eduskunnan alkuperäissäädöksiä suomeksi vuosilta 1734, 1868, 1889, 1895, 1896, 1898, 1901, 1906, 1907 ja 1917-2018.

[59] https://fi.wikipedia.org/wiki/MIT-lisenssi
[60] https://fi.wikipedia.org/wiki/Apache-lisenssi
[61] https://fi.wikipedia.org/wiki/GNU_General_Public_License
[62] https://fi.wikipedia.org/wiki/GNU_Lesser_General_Public_License
[63] https://www.kielipankki.fi/aineistot/
[64] http://urn.fi/urn:nbn:fi:lb-2019042612
[65] http://urn.fi/urn:nbn:fi:lb-2019042611

    - Aineisto on ladattavissa Kielipankin latauspalvelussa.
- Opusparcus: Open Subtitles Paraphrase Corpus for Six Languages (version 1.0)[66]
    - Opusparcus is a paraphrase corpus for six European languages: German, English, Finnish, French, Russian, and Swedish. The paraphrases are extracted from the OpenSubtitles2016 corpus, which contains subtitles from movies and TV shows.
- Finnish TreeBank 1[67]
    - 19,000 Sentences
    - 160,000 Tokens
- Suomen puupankki FinnTreeBank 2:n ladattava versio[68]
    - 163,197 Tokens
    - 19,197 Sentences
- Suomen puupankki FinnTreeBank 3:n ladattava versio[69]
    - 76,369,439 Tokens
    - 4,366,955 Sentences
- Kansalliskirjaston sanoma- ja aikakauslehtikokoelman OCR-korpus (1771 – 1874)[70]
    - ”Tämä korpus koostuu niiden Kansalliskirjaston digitoimien dokumenttien OCR-tuloksista, jotka on julkaistu ennen vuotta 1875.”
- Suomen kielen tekstikokoelman ladattava versio - kaupallinen käyttö[71]
    - Suomen kielen tekstikokoelma sisältää kirjoitettuja tekstejä 1990-luvulta muun muassa Helsingin Sanomista, Karjalaisesta, Kauppalehdestä, Tekniikan Maailmasta sekä WSOY:n kauno- ja tietokirjoista.

Jotkin julkiset organisaatiot tarjoavat kieliaineistoja omien avoimen datan jakamiseen tarkoitettujen verkkosivustojen kautta. Yksi esimerkki tallaisesta verkkopalvelusta on Eduskunnan Avoin Data[72], jossa kaikki kieliaineistot on jaettu JHS 189 -suosituksen[73] mukaisesti CC Nimeä 4.0 -lisenssillä[74]. JHS-suositus ehdottaa myös CC0-lisenssin käyttämistä silloin kun datan yksilöiminen ja datan tuottajan ilmoittaminen ei ole tarpeellista. Suomen julkisten toimijoiden dataa on kerätty erityisesti

[66] http://urn.fi/urn:nbn:fi:lb-2018021221
[67] http://urn.fi/urn:nbn:fi:lb-20140730138
[68] http://urn.fi/urn:nbn:fi:lb-2016042505
[69] http://urn.fi/urn:nbn:fi:lb-2016042601
[70] http://urn.fi/urn:nbn:fi:lb-2015051201
[71] http://urn.fi/urn:nbn:fi:lb-201908072
[72] http://avoindata.eduskunta.fi/index.html
[73] http://www.jhs-suositukset.fi/suomi/jhs189
[74] https://creativecommons.org/licenses/by/4.0/deed.fi

Väestörekisterikeskuksen ylläpitämään Avoindata.fi-palveluun[75], joka sisältää tällä hetkellä aineistoja yhteensä 802 organisaatiolta.

5.2. Puhekorpukset

Kielipankista löytyvät puhekorpukset kuten ”Eduskunnan täysistunnot, ladattava versio 1”[76] joka sisältää eduskunnan täysistuntojen äänitteet ajalta 10.9.2008 – 1.7.2016 sekä niiden kohdistetut transkriptiot, on lisensoitu ”CC-BY-NC-ND”. Tässä NC on lyhenne ”non-commercial” eli aineistoja ei saa käyttää yrityskäyttöön. Tämän ja muiden mahdollisten vastaavien aineistojen lisenssien uudelleenneuvottelu myös yrityskäyttöön olisi yksinkertainen tapa lisätä yrityksille käyttökelpoista materiaalia.

Tietosuojakysymysten vuoksi runsaasti henkilötietoa sisältävien puheaineistojen käyttöoikeuksien neuvotteleminen uudelleen ei useinkaan ole mahdollista. Tästä syystä on erityisen tärkeää kerätä uutta puheaineistoa kohdennetusti nimenomaan yrityskäyttöön. GDPR[77]-toimenpiteitä vaativia korpuksia, joiden lupien päivittämisen mahdollisuuksia kannattaisi selvittää, ovat muun muassa Sapu-korpus (n. 250 tuntia ääntä ja puhetta 2000-luvun Satakunnasta) ja Suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus (elisitoituja äänitekatkelmia sadoilta nykysuomalaisilta eri puolilta Suomea).

5.3. Avoimesti tarjolla olevat julkiset kieliresurssit

Suomenkielisiä kieliaineistoja syntyy osana julkishallinnon ja yritysten muuta toimintaa. Osan näistä kieliaineistoista keräävät talteen nämä toimijat itse, ja ne saattavat käyttää aineistoja sisäisesti esimerkiksi toimintansa kehittämiseen. Osalla näistä toimijoista on yhteistoimintaa kielentutkijoiden tai kieliaineistoja hyödyntävien yritysten kanssa, ja heidän kieliaineistojaan on koottu korpuksiksi tutkijoiden ja/tai yritysten käyttöön.

Joitakin huomioonotettavia julkisen sektorin toimijoita:

- yliopistot (erityisesti FIN-CLARIN-konsortion osapuolet ja Kielipankki) ja arkistot sekä Kotus
- Yle
- Eduskunta ja kunnat
- Kansaneläkelaitos, Verohallinto, Tilastokeskus ja Maahanmuuttovirasto

[75] https://www.avoindata.fi/fi
[76] http://urn.fi/urn:nbn:fi:lb-2017030901
[77] https://tietosuoja.fi/gdpr

Suomenkielisiksi kieliresursseiksi katsottavia palveluita on tarjolla myös julkisesti ylläpidettyinä. Esimerkiksi Kansalliskirjaston ylläpitämä Finto-ontologiapalvelu[78] on tällä hetkellä laajasti yritysten käytössä. Ontologiapalvelua voi käyttää hyödyksi esimerkiksi erilaisten hakujen rikastamisessa ja tulosten suodattamisessa. Esimerkiksi Lingsoft käytti Finto-palvelua Auria Biopankin rakentamisessa.[79] Mikäli Finto-palvelu loppuisi, olisi sellaiselle selvä tarve, joten nykyisen Finto-palvelun jatkuvuuden (ja erityisesti sen yrityskäyttöön soveltuvuuden) varmistaminen kuuluu osaltaan tämän ohjelman piiriin. Myös itse FIN-CLARIN konsortion ylläpitämä Kielipankki on tällainen kieliresurssi.

6. Suositukset

Yritysten kanssa käydyissä keskusteluissa on tullut selvästi ilmi, että yritysten tarpeet ja kyvykkyydet hyödyntää kieliteknologisia resursseja vaihtelevat hyvin paljon. Näissä toimenpidesuosituksissa on hahmoteltu operaatioita, jotka kokonaisuutena tuottaisivat arvoa suurimmalle osalle haastatelluista yrityksistä. Aineistojen ja ohjelmistokomponenttien luomisella, kokoamisella ja vapaasti tarjoamisella voi olla negatiivisiakin vaikutuksia niiden tahojen kannalta, jotka ovat jo käyttäneet paljon resursseja omien vastaavanlaisten aineistojensa ja ohjelmistojensa kehittämiseen, sillä nämä tahot menettävät näiden uusien resurssien myötä osan etulyöntiasemastaan. Uudet resurssit ovat kuitenkin myös näiden tahojen hyödynnettävissä, ja niillä on todennäköisesti uusia tulokkaita paremmat edellytykset resurssien käyttämiseen ja hyödyntämiseen omissa, jo valmiissa tuotteissaan. Jotkin kaavaillut ohjelmistokomponentteihin liittyvät operaatiot vievät esimerkiksi puheteknologian kehitystä varsin pitkälle. Niiden tarkoituksena on kuitenkin tuottaa vain peruskyvykkyyden mahdollistavat komponentit. Näiden peruskomponenttien jalostaminen, jatkokehittäminen ja yhdistäminen monimutkaisemmiksi kokonaisuuksiksi jättää alalla toimiville yrityksille vielä runsaasti toimintamahdollisuuksia.

Esitetyt operaatiot muodostavat tavallaan jatkumon, jossa ensimmäiset operaatiot luovat pohjaa myöhempien käytettäväksi. Tämä ei ole kuitenkaan näiden suositusten tarkoitus, ja operaatiot on suunniteltu erillään toteutettaviksi. Esimerkiksi tarkoitus ei ole täydellisen annotointialustan luominen ennen puheäänen korpuksen keräämistä. Puheaineistoa olisi syytä ryhtyä keräämään annotointialustaprojektin olemassaolosta riippumatta niillä välineillä, joita tällä hetkellä on käytettävissä tai kyseisen operaation resursseilla tuotettavissa. Toki yhteistyö operaatioiden välillä on toivottavaa ja suotuisaa. Jokaisen operaation voisi mahdollisuuksien mukaan aloittaa työpajalla, jossa tarkemmin kehitettäisiin ja määriteltäisiin kyseisen operaation tavoitteita yhdessä

[78] https://finto.fi/fi/
[79] https://www.lingsoft.fi/en/asiakkaat/terveydenhuolto/mining-for-gold-patient-records

kiinnostuneiden tahojen kanssa. Erityisen tärkeää on, ensimmäisen operaatioaihion mukaisesti, varmistaa sellaisen tahon olemassaolo, joka huolehtii muodostettavien suomalaisten kieliresurssien ylläpidosta, kehityksestä, ja tarjonnasta.

Puheaineiston litterointitapa ja -muoto vaikuttaa oleellisesti aineiston käyttökelpoisuuteen. Suurimpaan osaan luonnollisia puhetilanteita liittyy usean puhujan välinen vuorovaikutus. Esimerkiksi eri puhujien puheen ajallinen limittyminen on huomioitava litteroinnissa ja käytettävä puheelle sopivaa litterointimenetelmää, joka sallii puhujien erottamisen omiin litterointikerroksiinsa ja siten esimerkiksi päällekkäispuhunnan esittämisen ilman turhia erikoismerkkejä. Koneoppimisen kannalta on keskeistä, että koko puheaineisto litteroidaan systemaattisesti ja yhdenmukaisella tavalla. Sanallisen sisällön lisäksi on merkittävä tarkoituksenmukaisella tavalla erikseen esimerkiksi täytesanat, kesken jääneet sanat ja muut äännähdykset. Automaattisen jälkikäsittelyn vuoksi olisi parasta, että suurin osa litteraatioista olisi tallennettu käyttäen XML-pohjaista formaattia. Tällöin aineistosta on helppo jättää huomiotta opeteltavan tehtävän kannalta tarpeeton annotaatio.

Tämän hankekokonaisuuden kautta hankituissa tai luoduissa kieliaineistoissa olisi syytä ottaa kattavasti huomioon kaikenlainen suomen kielen vaihtelu, niin tyylilajien ja murteiden kuin puhujien syntyperän, terveydentilan tai iän mukaan. Erityisesti olisi myös syytä huomioida maahanmuuttajien ääntämiseen ja kielitaitoon liittyvät erityispiirteet. Voidaan ajatella, että eri maahanmuuttajaryhmät puhuvat suomea käyttäen eräänlaisia uusia murteita.

Suomen kirjakielen kirjoitusjärjestelmää voi pitää melko fonemaattisena. Puheessa on useita ilmiöitä, joita ei merkitä näkyviin mitenkään normaalissa kirjoitetussa tekstissä.[80] Yleiskielen ortografiaa myötäilevä litterointi riittää monessa tapauksessa puheen sisällön kuvaamiseen, mutta tarkempaa litterointia tarvitaan, jos puheen tekstimuotoisessa kuvauksessa halutaan tarkemmin esittää esimerkiksi puhujien taustan alueellisia, sosiaalisia ja tilanteisia eroja.

Automaattisessa puheentunnistuksessa tavoitteena on yleensä ollut kuvata puheen sisältö pitkälti ”normaalikirjoitusta” myötäilevällä litterointitavalla. Lukuisiin käyttökohteisiin tällainen lopputulos riittääkin varsin hyvin. Sovellusten merkittävä kehitys kuitenkin edellyttää, että myös puheen hienojakoisempia ja kirjoittamattomia ominaisuuksia voidaan jatkossa hyödyntää paremmin.

[80] Esimerkiksi assimilaatiot, äng-äänne, diftongien redusoituminen tai avartuminen ja sanapaino.

Kielitieteessä on kehitetty erilaisia foneettisia tarkekirjoitusjärjestelmiä lähinnä siksi, että puhenäytteiden tarkempia ääntеellisiä ja muita ominaisuuksia voitaisiin saattaa lukukelpoiseen ja helpommin tutkittavaan muotoon. Foneettisen kirjoituksen kultakaudella digitaalisten ääni- tai videonäytteiden tallentaminen ja tutkiminen ei ollut mahdollista tai ainakaan yhtä helppoa ja nopeaa kuin nykyisin, jolloin puheen muistiin merkitseminen oli tärkeää. IPA-merkistö[81] ja siihen nojaava foneettinen transkriptio ovat edelleen laajasti käytössä maailmalla, minkä lisäksi on eri kieliryhmiä tai kieliä koskevia transkriptio-merkistöjä ja litterointitapoja. Esimerkiksi suomalais-ugrilaisten kielten kuvausta varten on kehitetty suomalais-ugrilainen transkriptiojärjestelmä eli SUT[82], josta ei kuitenkaan ole tällä hetkellä tarjolla yhtenäistä kuvausta. Foneettinen kirjoitus ei välttämättä esitystapana ole puheteknologian kannalta tehokas lähestymistapa, koska sen tulos on jossain määrin subjektiivinen ja tarkekirjoituksen tuottaminen ihmisvoimin on hidasta ja kallista. Jotkut foneettiseen kirjoitukseen sisältyvät ajatukset voisivat kenties kuitenkin olla myös koneoppimisen kannalta hyödyllisiä, esimerkiksi äänteiden kuvaaminen niiden artikulaatioon ja muuhun puhe-elimistön toimintaan liittyvien piirteiden mukaisesti pelkkien kirjainsymbolien sijaan. Toisaalta esimerkiksi puheentunnistuksen lopputuotteena halutaan kuitenkin useimmiten kirjakielen kaltaista tekstiä, joka pitäisi sitten jälkikäteen generoida järjestelmän sisäisen esitysmuodon perusteella.

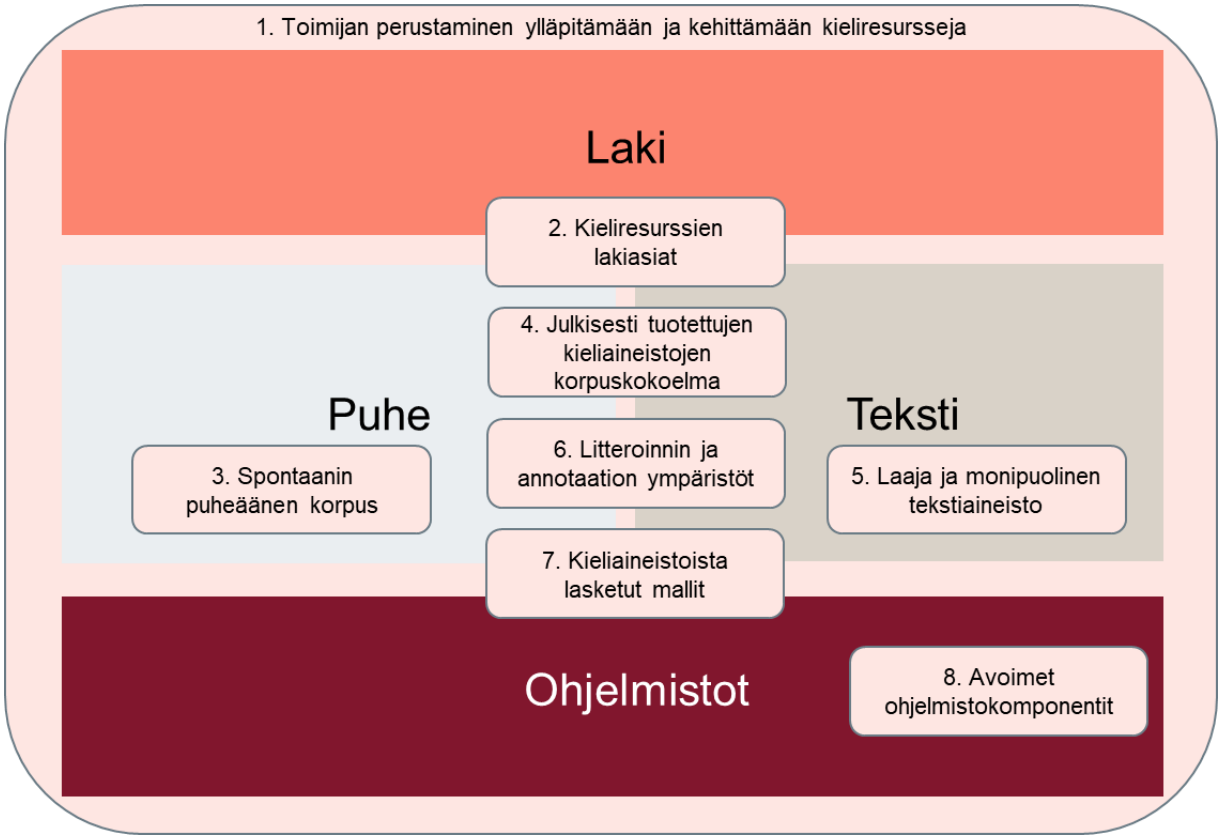
Puheteknologian osalta tulisikin selvittää, millainen puheen litterointi- ja transkriptiomenetelmä (tai muu systemaattinen puheen ominaisuuksien kuvaustapa) olisi järjestelmän toiminnan kannalta optimaalinen. Tällaisten yhtenäisten ”litteraatiokäytänteiden” määritteleminen ja julkaiseminen helpottaisi myös eri puheaineistojen yhteiskäyttöä tulevaisuudessa, koska kaikki aineistot voitaisiin pyrkiä kuvaamaan vähintään puheteknologian kannalta hyödyllisellä ja yhteensopivalla minimitasolla.

Tekstiaineiston annotaatiossa ei ole myöskään tarjolla selvää standardia, mutta tämän hankekokonaisuuden aineistoissa annotaation olisi hyvä olla yhtenäistä. Kielipankissa suositellaan käytettäväksi tekstiaineistoihin XML-pohjaista formaattia, joista laajiten käytetty standardi on TEI[83]. TEI:tä tukevista editoreista löytyy lista osoitteesta: https://wiki.tei-c.org/index.php/Editors

Suomenkielisen tekoälyn kehittämisohjelman operaatioita esitellään tässä esiselvityksessä yhteensä kahdeksan. Panostuksesta riippuen ne olisi mahdollista

[81] International Phonetic Alphabet: https://en.wikipedia.org/wiki/International_Phonetic_Alphabet
[82] IPA- ja SUT-tarkekirjoitusjärjestelmät on kuvattu mm. julkaisussa Iivonen, Sovijärvi & Aulanko (1990) Foneettisen kirjoituksen kehitys ja nykytila, Helsingin yliopiston fonetiikan laitoksen monisteita nro 16.
[83] Text Encoding Initiative: https://tei-c.org

toteuttaa täydessä laajuudessaan vuoden 2022 loppuun mennessä. Operaatiot toteutetaan kullekin tavoitteelle sopivimmalla tavalla, osin käytetään hyödyksi jo olemassa olevaa yhteistyötä, osin luodaan uusia kumppanuussuhteita.



Kuva 1. Suunnitellut operaatiot suhteessa puhe- ja tekstiaineistoihin, lakiasioihin ja ohjelmistoihin.

6.1. Organisaatio suomalaisten kieliresurssien kehittämiseen ja ylläpitoon

Tässä hankekokonaisuudessa kehitettävät kieliresurssit tarvitsevat toimijan, joka varmistaa, että kieliresurssit ovat pitkäaikaisesti ja luotettavasti saatavilla. Tavoitteena on operaatioiden lopputulosten ja niiden tuottamien prosessien jatkuvuuden sekä saatavuuden varmistaminen. Tätä varten tarvitaan organisaatio, joka huolehtii aineistoista, ohjelmistoista ja malleista, sekä ylläpitää ohjeistoa siitä, miten näitä resursseja (”assets”) käytetään, jaetaan, ja hyödynnetään. Sama organisaatio toimisi yhteistyötahona kaikkien suomenkielisestä kieliteknologista kiinnostuneiden tahojen välillä (käyttävät ja toimittavat tahot sekä tutkijat).

Toimijan olisi syytä olla neutraali suhteessa alalla toimiviin yrityksiin ja näiden tulisi voida luottaa siihen, että se myös pysyy sellaisena. Kanavia kieliresurssien jakeluun voisi olla myös useita, mutta perustettavan organisaation olisi joka tapauksessa syytä koordinoida nimenomaan suomenkielisiä kieliresursseja kokonaisuutena. Osa aineistoista tai ohjelmistoista voisi olla esimerkiksi tulevaisuudessa saatavissa ensisijaisesti FIN-CLARIN konsortion ylläpitämän Kielipankin tai esimerkiksi Euroopan laajuisen ELG:n[84] kautta. ELG:n kieliresurssien jakelualustan pitäisi valmistua vuoden 2021 lopulla.

Ajatuksena on, että organisaatio koordinoi kieliresurssien ympärillä tapahtuvaa kehitystä, mutta ei välttämättä tee kaikkea itse. Se voisi esimerkiksi toimia osapuolena ulkopuolisen, esimerkiksi julkisen, rahoituksen hankkeissa, joissa kieliresursseja edelleen kehitetään johonkin tiettyyn käyttötarkoitukseen sopiviksi tai luodaan kokonaan uusia resursseja. Organisaation tulee toimia läheisessä yhteistyössä akateemisen puolen kieliresursseja vastaavasti keräävän ja kehittävän FIN-CLARIN-konsortion ylläpitämän Kielipankin kanssa. Mahdollisuuksien mukaan kieliresurssit tarjotaan käyttöön ja löydettäväksi samojen työkalujen avulla kuin Kielipankin tutkijoille tarkoitetut resurssit. Kaikki resurssit tulisi kuvailla CSC:n ylläpitämään META-SHARE-palveluun[85], joka perustuu eurooppalaisessa META-NET-hankkeessa rakennettuun verkostoon. Kieliresurssit voivat sijaita hajautetusti, kunhan niiden jatkuva saatavuus on turvattu kullekin resurssille sopivalla varmuudella.

Vaikka tässä kehittämisohjelmassa onkin keskitytty suomenkielisten kieliresurssien kehittämiseen, ei ole tarkoituksenmukaista rajata organisaation toimintaa siihen. Sopiva määritelmä voisi olla esimerkiksi ”Suomessa käytettävien kielten kielivarojen edistäminen”, ja – esimerkiksi säätiömuotoiselle toimijalle – nimiehdotus voisikin kuulua ”Suomalaisten kielivarojen kehittämissäätiö” [86], ”Foundation for the advancement of language resources in Finland”.

Kieliresurssien käyttöä edistetään mm. tiedottamalla niiden olemassaolosta mahdollisimman laajasti käyttäen sellaisia kanavia, joita kieliresurssien potentiaalinen käyttäjäkunta seuraa. Yksi tärkeä tapa edistää resurssien käyttöä ja varmistaa resurssien käyttökelpoisuus on säännöllinen yhteydenpito kaikkiin toimijoihin, joilla on kiinnostusta resursseihin.

[84] https://www.european-language-grid.eu/about/
[85] http://metashare.csc.fi
[86] Kielivara on kieliresurssi-sanan synonyymi, jota nykyään käytetään FIN-CLARIN yhteyksissä, esimerkiksi: https://www.eduskunta.fi/FI/vaski/JulkaisuMetatieto/Documents/EDK-2015-AK-11433.pdf

Suuri osa suomenkielisestä kieliteknologiasta perustuu yliopistoissa tehtyyn kieliteknologian ja kielitieteiden perustutkimukseen. Organisaation perustamista edistetään suomenkielisten kieliresurssien kehittämiseen kohdistetulla operaatiolla, jossa tehdään konkreettinen ehdotus siitä, miten organisaatio tuottaisi pitkäaikaista hyötyä yrityksille ja yhteiskunnalle.

Tässä operaatiossa etsitään vastaukset mm. seuraaviin kysymyksiin ja lähdetään toteuttamaan operaation toimivallan puitteissa työtä käytännössä:

- Miten varmistetaan jakelukanavan pitkäjänteinen toiminta ja saavutettavuus? Minkälainen toiminta-, liiketoiminta- sekä rahoitusmalli olisi mahdollinen?
- Onko sama toimija niin tutkimuksen kuin yritysmaailman resurssi?
- Olisiko käytännöllistä esim. ylläpitää portaalia, jossa listataan kieliaineistoista kiinnostuneita yrityksiä ja niiden pohjalta tai niihin liittyviä palveluita tarjoavia yrityksiä?
    - Ylemmän tason toimijat kaipaavat tietokantaa palveluntarjoajista suomenkieliseen tekoälyyn liittyen.
    - Näin pyritään muodostamaan kieliresurssien kehittäjien ja niiden tarvitsijoiden välille verkosto nopeuttamaan ja mahdollistamaan sopivien yhteistyökumppanien löytämistä.
- Huolehtisiko tämä organisaatio joltain osin myös Suomessa järjestettävään suomenkieliseen kieliteknologiaan ja tekoälyyn liittyvästä koulutuksesta ja koulutusmateriaalista?
- Tekisikö organisaatio, puoluettomana toimijana, eri kieliresurssien alueella toimivien yritysten tuotteiden vertailuanalyysia[87]?
- Voiko organisaatio toimia kokoavana yhteistyötahona haettaessa rahoitusta tutkimus- ja tuotekehityshankkeisiin (sekä yliopistot että yritykset)?

6.2. Kieliresurssien sekä perustettavan tai toimintaansa laajentavan organisaation lakiasiat

Lakiasioiden selvittäminen sekä perustettavan tai toimintaansa laajentavan organisaation että kieliresurssien keräämisen ja hyödyntämisen näkökulmasta on syytä käynnistää mahdollisimman nopeasti.

Ennen minkäänlaisten teksti- tai puheaineistojen keräämistä on syytä selvittää tarkemmin niitä koskeva sääntely erityisesti tekijänoikeuden ja tietosuojan osalta, ja laatia kattava, mutta selkeä ohjeistus. Erityisesti tekstiaineistojen ja luetun tekstin

[87] https://www.kielikello.fi/-/benchmarking

kohdalla on huomioitava tekijänoikeuskysymykset ja näihin liittyen esimerkiksi kopioiden tekeminen ja edelleen luovutus. Vastaavasti äänitietoa on henkilötiedon laajan määritelmän vuoksi turvallisinta käsitellä henkilötietona; tällöin on huomioitava esimerkiksi yleinen tietosuoja-asetus (GDPR) ja minimoitava käsiteltävät tiedot ja identifioivat tekijät. Lisäksi on joissakin tapauksissa huomioitava myös muuta sääntelyä, joka liittyy esimerkiksi toimialaan (esimerkkinä terveydenhuolto), julkisuus- tai virkamieslakiin, sopimuksiin (esim. salassapito), viestinnän sääntelyyn, ym. Näin varmistetaan, että aineistojen käyttö ja luovutus myös kaupallisessa tarkoituksessa ja riittävässä laajuudessa on mahdollista: On huomioitava, ettei käyttäjälle voida antaa enempää oikeuksia kuin luovuttajalta on saatu, eli ketjun on oltava ehjä.

”Kieliresurssien sekä perustettavan tai toimintaansa laajentavan organisaation lakiasiat” -operaation kohteena ovat aineistojen keräämiseen, käyttämiseen ja jakamiseen liittyvät tietosuoja- ja tekijänoikeuskysymykset. Työn tarkoituksena on tuottaa yleinen ohjeistus lakikysymyksiin, jotka liittyvät kieliresursseihin, niiden keräämiseen ja jakamiseen, sekä tuottaa erilaisia yleisten tason sopimusmalleja kehitystyötä ja aineistojen luontia sekä hyödyntämistä helpottamaan. Monilla, varsinkin pienemmillä, teknologiayrityksillä on rajoittuneet mahdollisuudet konsultoida lakimiehiä kerätessään ja käsitellessään kieliaineistoja. Esimerkiksi käsittelemättömiin puheaineistoihin tulisi kuitenkin soveltaa GDPR-säännöksiä, koska ihmisen luonnollisen puheen katsotaan olevan ihmisen yksilöivä henkilötieto.

Minkälaiset sopimukset puhujien kanssa riittävät, jotta puheen käsittely yrityksen sisällä on säädösten mukaista? Tekoälykiihdyttämön (FAIA) tuottama AI playbook[88] käsittelee aihetta yleisellä tasolla, mutta yrityksillä on tarve selvästi pidemmälle vietyihin mallisopimuksiin ja muuhun tarkempaan ohjeistukseen. Milloin puhe- tai tekstiaineistolla on tekijänoikeudet? Voiko yritys turvallisesti käyttää CC0[89]-lisensioitua aineistoa taholta x? Mitä GDPR edellyttää? Näihin ja moniin muihin lakiteknisiin kysymyksiin olisi syytä luoda yksityiskohtainen ohjeistus mallisopimuksineen, ja tätä ohjeistusta tulisi myös ylläpitää, koska käytännöt ja käyttötapaukset muuttuvat vuodesta toiseen.

Operaation tehtävänä on yhteistyössä tähän osallistuvien yritysten kanssa kerätä yhteen erilaiset käyttötapaukset, joissa erilaisia aineistoja kerätään ja jaetaan niitä käyttäville tahoille. Näiden aitojen tarkkojen käyttötapausten perusteella hahmotellaan yleisimpiä skenaarioita, joissa lakikysymyksiä joudutaan ratkomaan suhteessa kieliresursseihin. Näiden yleistettyjen tapausten ratkaisemiseksi kirjoitetaan ohjeistus ja luodaan niitä varten tarvittavia mallisopimuspohjia, joita on mahdollista käyttää pohjana vastaavan kaltaisissa tilanteissa.

[88] https://faia.fi/playbook/
[89] https://creativecommons.org/publicdomain/zero/1.0/

Operaatioon kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Ohjeistus ja mallisopimuksia, joissa GDPR ja tekijänoikeuskysymykset on käsiteltynä niin pitkälle kuin mahdollista, jotta yritysten on yksinkertaista muokata mallisopimuksista omaan tarkempaan tarpeeseensa sopiva sopimus
    - Ohjedokumentti, joka selittää pohjien käytön ja yleisellä (tai tarkallakin) tasolla käy läpi erilaisiin kieliresursseihin liittyvät ongelmat
    - GDPR:n alaisen aineiston keräämiseen ja jakamiseen sopimuspohjat
    - Avointen lisenssien käyttö
- Patentit
- Pohjasopimus helpottamaan satunnaisen verkkojulkaisijan tekemiä aineistoluovutuksia

Mitä tulee perustettavan tai toimintaansa laajentavan organisaation lakiasioihin, on tarkoituksena selvittää ja mahdollisen investointipäätöksen liittyen organisaatioon ja/tai sen perustamiseen jälkeen toteuttaa malli (juridinen, modus operandi) joka toteuttaa parhaiten organisaatiolle asetettuja tavoitteita ja mahdollistaa suunnitellun (liike)toimintamallin toteutumisen.

6.3. Spontaanin puheäänen korpus

Kolmas operaatio on laaja yleiskäyttöinen suomenkielinen puhekorpus. Puheen tulisi olla mahdollisimman luonnollista spontaania puhetta. Pelkästään tietoisuus puheen tallentamisesta muuttaa puhujien rekisteriä[90] ja puhujien puhetapaa. Tämän vuoksi ei esimerkiksi kannattaisi erityisesti painottaa keräävänsä tiettyjä murteita vaan pikemminkin pyrkiä saamaan puhuja unohtamaan, että hänen puheensa tallennetaan, jolloin on mahdollista kerätä luontevampaa kieltä.

Aikaisempien laajojen puheaineistojen keruuhankkeiden kokemuksia olisi syytä ottaa huomioon uusia projekteja suunniteltaessa ja mahdollisuuksien mukaan käyttää näissä hankkeissa kehittynyttä alan asiantuntemusta. Yksi huomioitavista hankkeista on Koneen Säätiön rahoittama, vuonna 2013 toimintansa aloittanut Prosovar-hanke[91], jossa tavoitteina oli ”tutkia puhesuomen kielen prosodisia piirteitä ja siinä ilmenevää alueellista ja sosiaalista variaatiota, muodostaa erityisesti prosodian ja sen variaation tutkimukseen soveltuva puhesuomen korpus sekä kehittää ja testata uusia,

[90] https://tieteentermipankki.fi/wiki/Kielitiede:rekisteri
[91] Sivu 56: http://urn.fi/URN:ISBN:978-951-29-5980-8

vähintäänkin osin joukkoistettuja menetelmiä puheaineistojen keräämiseksi ja tutkimiseksi internetin välityksellä.”[92]

Korpus lisensoitaisiin mahdollisimman avoimesti ja sitä jaettaisiin niin pienellä byrokratialla kuin GDPR:n ym. puolesta on mahdollista. Monet haastatellut puheteknologiaa hyödyntävät tai kehittävät yritykset korostivat tarvetta nimenomaan oman tarkan kohdealueensa aineistoille, jota käytännössä kannattaa kerätä suoraan kehitettävistä palveluista itsestään. Yleisempää suomea ymmärtävälle puheentunnistimellekin on kuitenkin merkittävästi käyttökohteita esimerkiksi sellaisissa palveluissa, joissa tekstiksi muutettua puhetta käytetään lähinnä tiedonhakuun, tai sellaisissa, joissa tarkasti määriteltyä puheen aihealuetta ei alun alkaenkaan ole.

Luetun tekstin ja vapaan puheen ero on myös tullut haastattelujen aikana selväksi. Käytännön palveluissa puheentunnistimen tarkoitus on lähes aina tunnistaa nimenomaan vapaamuotoista, ei luettua, puhetta. Joissakin käyttötapauksissa, esimerkiksi lääkärien saneltujen muistiinpanojen muuntamisessa tekstimuotoon, puhetyylinä on kuitenkin ääneen lukeminen. Myös sopivasti suunnitellun, äänneyhdistelmiltään tasapainotetun ääneen luetun materiaalin keräämiselle voi olla perusteita, joskin suurin pula on juuri aidosta keskustelupuheesta. Tällaista materiaalia voitaisiin kerätä vapaamman aineiston keruun yhteydessä esimerkiksi alkuverryttelynä vapaamuotoiseen keskusteluun, mikäli tällainen saataisiin sopimaan keräystilanteeseen sujuvasti.

Tarkoituksenmukaista olisi pyrkiä tekemään yleisestä puhekorpuksesta mahdollisimman laaja niin, että se sisältäisi erilaisia aihealueita ja erilaisia puhujia (ikä, sukupuoli, murretausta). Kaikki aihealueiden ja puhujien relevantit taustatiedot on merkittävä mukaan korpukseen tarkasti, jotta siitä olisi tarpeen mukaan irrotettavissa myös erilaisia osakorpuksia, jos vaikkapa halutaan tuottaa palvelua jollekin tietylle murrealueelle.

Tasapainon löytäminen projektin laajuuden suhteen heti alussa on tärkeää. Laajuus riippuu erityisesti käytettävissä olevista rahallisista resursseista ja projektin käytettävissä olevista asiantuntijoista (yritykset ja akateemiset tutkijat) sekä mahdollisista promoottoreista (esim. Yle). Tällaisen avoimesti saatavissa olevan, tarkasti litteroidun puheaineiston muodostaminen ja saataville asettaminen olisi hyödyksi käytännössä kaikille puheteknologiaa kehittäville tahoille.

Operaatioon kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

[92] Sivulta 29 alkaen: http://fp2015.aalto.fi/Fonetiikan_Paivat-2015_Aalto-yliopisto.pdf

- Yhteistyö esim. Ylen kanssa (Yle promoottorina) puheaineistojen keräämiseksi
    - Kerättyjen aineistojen käyttöehdoista pitää sopia yhteisesti ennen projektin aloittamista Yle-vetoisena
    - Minkälaiset ohjelmatuotannot voisivat sopia tällaiseen tarkoitukseen?
- Asetetaan tavoite korkealle: 10 000 tuntia puhetta, 5-15 minuuttia per puhuja, eli noin 40-100k puhujaa (tämä mahdollista vain Ylen vahvalla tuella)
    - Puhujien taustatiedoista pitäisi mahdollisuuksien mukaan tallentaa:
        - Ikä, sukupuoli, murretausta (mitä murretta mielestään puhuu), onko suomi 1. kieli, äidinkieli
    - Tavoitteena saada puhujat keskustelemaan eri aihealueista
    - Luetutettaisiin puhujilla ensin 10-20 lausetta etukäteen valmistellusta materiaalista
    - Käytetään aineistonkeruun asiantuntijoita
- Litteroidaan ja annotoidaan laadukkaasti, pyritään löytämään tarkoituksenmukaiset käytännöt
    - Ei ole itsestään selvää mitä litterointitarkkuutta tai järjestelmää puhetta litteroidessa tulisi käyttää
    - Valittavalta litteroijataholta olisi syytä pyytää työnäyte
        - Työnäytteestä selviää litteroinnin laatu ja kuinka paljon työhön kuluu aikaa
    - Litterointiin olisi syytä alustavasti varata noin 150 – 250 € jokaista nauhoitettua äänitetuntia kohden (tarkka hinta riippuu litteroinnin tarkkuudesta ja litteroitavasta materiaalista)
- Käytännössä litteroimatonta ja annotoimatontakin puhetta varmasti tarvitaan, joten kerättävän puheen määrää ei kannata rajata vain sen perusteella, että pelätään, ettei sitä saada litteroitua ja annotoitua käytössä olevilla resursseilla

6.4. Julkisesti tuotettujen kieliaineistojen korpuskokoelma

Neljäntenä operaationa olisivat helpommin hankittavat, rakennettavat tai jo olemassa olevat kieliaineistot. Tarkoitus olisi luoda olemassa olevista ja “automaattisesti” karttuvista julkisesti tuotetuista kieliaineistoista korpuskokoelma yrityskäyttöön (sekä puhe että teksti).

Kohdan 6.3 mukaisen puhekorpuksen lisäksi olisi hyvä mahdollisuuksien mukaan kerätä myös vähemmillä resursseilla käyttöön asetettavia puhekorpuksia. Myös eräiden julkisten tahojen automaattisesti tuottamien karttuvien aineistojen ympärille olisi hyvä luoda ylläpidetty järjestelmä, jonka prosessien kautta aineistot tulisivat yleiseen käyttöön myös yrityksille.

Yliopistoissa ja korkeakouluissa on kerätty erilaisia puheaineistoja, mutta monen tällaisen aineiston käyttö on rajattu eksplisiittisesti tai implisiittisesti tieteelliseen tutkimukseen ja opetukseen. FIN-CLARIN-konsortion ylläpitämästä Kielipankista[93] löytyy tällä hetkellä monipuolisesti erilaisia teksti- ja puheaineistoja tutkijoiden käyttöön. Osa näistä aineistoista on käytettävissä myös yrityskäyttöön. Myös aineistojen käyttö tai luovutus ns. kolmansille osapuolille voi olla eksplisiittisesti tai implisiittisesti rajattu. Koska tällaisia aineistoja kuitenkin on, olisi järkevää pyrkiä selvittämään mahdollisuudet tarvittavien lupien hankkimiseen jälkikäteen. Tutkijoiden käytössä jo olevien kieliaineistojen vapaamman jakelun mahdollisuudet olisi selvitettävä. On kuitenkin todennäköistä, että jo kerättyjen puheaineistojen puhujia on vaikea enää tavoittaa.

Operaatioon kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Jo valmiiksi sopivasti lisensoitujen ja Kielipankista löytyvien aineistojen identifiointi
    - GDPR:n alle kuuluvien puheaineistojen osalta pitäisi pyytää uudet luvat myös puhujilta itseltään.
- Eduskunnan ja kunnanvaltuustojen istuntojen puheenvuorot ja keskustelut
    - Osin valmiiksi sekä puheena että tekstinä
    - Tommi Kurki ja Camilla Wide Turun yliopistosta ovat suunnittelemassa kaupunginvaltuustojen kokoustallenteista muodostuvaa korpusta
        - Näissä tallenteissa tulevat paikalliset (murre) erot paremmin esille kuin eduskunta-aineistossa
        - Äänenlaatu on heikko osassa tallenteita, koska kokous on saatettu tallentaa esimerkiksi yhden paikallaan pysyvän videokameran sisäisellä mikrofonilla
- Sopimusneuvottelut
    - Lisenssiehdoista pitäisi erityisesti saada pois NC (non-commercial) kohdat
- Annotointi, puheen litterointi
    - yhtenäinen ohjeistus litterointiin
- Aineistoehdotuksia
    - Julkisten tahojen kyselyt, palautepyynnöt, yms. ja niihin tulleet vastaukset

6.5. Laaja ja monipuolinen tekstiaineisto

Useilla toimijoilla on jo käytössään laajoja suomenkielisiä tekstiaineistoja, joko suoraan omilta asiakkailta saatuja tai itse verkosta kerättyjä.[94] Tarve onkin ennen kaikkea isolle,

[93] https://www.kielipankki.fi

[94] Esimerkiksi suomenkielinen Wikipedia on ladattavissa yhtenä pakettina, jonka jälkeen siitä voi laskea erilaisia malleja.

monipuoliselle aineistolle, johon on laajasti annotoitu aineistoa ja sen käyttämää kieltä koskevaa metatietoa, kuten sentimenttejä, murretietoa tai aihealueita.

Viidentenä operaationa on ison suomenkielisen tekstiaineiston hankinta. Tavoitteena on hankkia tekstiaineisto kokonaan avoimeen käyttöön, tai yhteistyössä aineiston omistajan kanssa suunnitella ja neuvotella ketterästi toimiva prosessi aineiston hankintaan. Tarkoitus on saada monipuolinen pohja erilaisten suomenkielisessä tekstissä esiintyvien ilmiöiden mallien laskentaan ja tutkimiseen. Aineiston kielen tulisi olla mahdollisimman luonnollista ja monipuolista, mikä tarkoittaa, että kielen tulisi olla myös muuta kuin virka- tai toimistotyönä (esim. julkishallinnon www-sivustot) tuotettua tekstiä tai tietosanakirjamaista tekstiä (esim. Wikipedia). Myös kirjoitetun kielen vaihtelua esimerkiksi murteiden käytön osalta olisi syytä sisältyä aineistoon. Vastaavanlaisia aineistoja on tarjolla myös kaupallisesti (esim. www.futusome.com), mutta näiden toimijoiden toimintamallit ovat yleensä laajempia ja reaaliaikaisia. Laajana pohja-aineistona hankittavan tekstiaineiston ei tarvitsisi olla reaaliaikainen (vaikkakin karttuva), vaan se voisi esimerkiksi olla jonkin keskustelufoorumin kaikki viestit edellisen vuoden loppuun asti, eli tätä esiselvitystä kirjoittaessa vuoden 2018 loppuun.

Suomi24 on tällä hetkellä suurin tällainen aineisto Kielipankissa ja se sinällään sopisi tähän tarkoitukseen,[95] mutta vastaavien aineistojen tarkempi kartoittaminen projektin aluksi on järkevää. Myös ylilauta.org-palvelun tekstejä on tällä hetkellä saatavissa Kielipankin kautta tutkimuskäyttöön.[96] Myös muiden isojen alkuaan erikoistuneiden keskustelupalstojen, kuten hevostalli.net ja vauva.fi, keskustelujen aihealueet ovat nykyisin laidasta laitaan.

Suomi24:n tai muun valitun aineiston oikeuksien omistajan kanssa pitäisi neuvotella mahdollisuudesta saattaa koko tekstiaineisto kaikkien yritysten käyttöön avoimella lisenssillä. Mikäli täysin avoimen lisenssin hankkiminen osoittautuu liian kalliiksi tai muuten haastavaksi, toinen vaihtoehto on neuvotella yleinen kaava aineiston hinnalle, jolla kukin yritys voisi sitten aineistoa tarvitessaan sen hankkia. Tällöinkin aineiston jakelukanavan toiminnan ja pysyvyyden varmistaminen kuuluisi operaation tehtäviin.

Operaatiossa aineisto annotoidaan mahdollisimman laajasti (morfologia, syntaksi, nimetyt entiteetit, sentimentit, tunnelataukset) manuaalisesti, puoliautomaattisesti ja automaattisesti. Annotoinnin määrä ja taso määräytyvät projektiin käytettävissä olevien resurssien perusteella. Annotaatiot olisivat saatavissa erikseen avoimella lisenssillä ja niiden pitäisi olla yhdistettävissä erillisesti jaettavaan aineistoon.

[95] http://urn.fi/urn:nbn:fi:lb-2019010801
[96] http://urn.fi/urn:nbn:fi:lb-2016101210

Tähän kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Vaihtoehtoja:
    - Suomi24.fi
    - Muita: ylilauta.org, vauva.fi, forum.hevostalli.net, punkinfinland
- Sopimusneuvottelut
- Aineiston annotointi, annotoidaanko kaikki kaikilla tasoilla vai pitääkö tehdä kompromisseja (kustannuskysymys)
    - Morfologia
    - Syntaksi[97]
    - NER
    - Sentimentit
    - Tunnelataukset (viha, ilo, yllätys, inho, häpeä, pelko)

6.6. Kieliaineistojen litteroinnin ja annotaation ympäristöt

Kuudentena operaationa ovat kieliaineistojen keräämiseen ja erityisesti niiden annotointiin tarkoitetut ympäristöt. Haastatteluissa on käynyt selväksi, että toimiakseen mahdollisimman hyvin mikä tahansa kieltä käyttävä järjestelmä kannattaa ainakin osin opettaa aineistoilla, jotka ovat hyvin lähellä kutakin yksityiskohtaista käyttötapausta. Käyttötapausten monimuotoisuuden ja uusien käyttötapausten jatkuvan syntymisen vuoksi on mahdotonta tuottaa yleisiä harjoitusaineistoja kaikkiin tarpeisiin. Tämä johtaa jatkuvaan ja pysyvään kieliaineistojen annotointitarpeeseen.

Sekä litteroinnissa (transkriptiossa) että annotoinnissa on tällä hetkellä vaikea määritellä, mitä tiettyä ympäristöä tai sovellusta kannattaisi käyttää tai miten litterointi tai annotointi pitäisi tehdä. Suositeltavan litteraation ja annotoinnin tarkkuus riippuu yleensä aineiston ja sen avulla tehtävän ohjelmiston tai palvelun käyttökohteesta. Tutkimuskäyttöön tarkoitettu annotointiympäristö ei välttämättä ole optimaalinen yritysten tarpeisiin.

Yritysten kannalta olisi hyvä, mikäli olisi käytettävissä yleisesti ylläpidetty avoimen lähdekoodin web-käyttöliittymällinen ohjelmisto, jolla olisi mahdollista tuottaa standardimuotoista litteraatiota ja annotaatiota eri tasoilla. Yritykset voisivat joko itse asentaa ohjelmiston sisäiseen suljettuun käyttöönsä tai ne voisivat ostaa sen ylläpidon, tai jokin yritys voisi myydä ohjelmistoa esimerkiksi alustapalveluna. Myös yhteisesti ylläpidetylle annotointiympäristölle voisi olla kysyntää. Annotointiympäristö olisi hyvä

[97] Aineistosta pitää annotoida käsin monipuolinen otos, jolla testataan nykyisten automaattisten morfologisten ja syntaktisten analysaattoreiden taso. Testauksen tulosten perusteella päätetään, onko taso riittävä.

yhdistää ekosysteemiin, jonka kautta voisi helposti rekrytoida erilaisia annotoijia (vaikkapa joukkoistaa annotointityötä) mahdollisimman virtaviivaisesti.

Tähän operaatioon kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Annotaatioalustojen ja annotaatiosuositusten koonti
    - Mitä potentiaalisesti käyttökelpoisia annotaatioalustoja tai käytäntöjä on jo olemassa?
        - Olisiko annotointialustan syytä olla selainpohjainen?
        - Tutkijat käyttävät yleisesti erillisesti omalle työasemalle asennettavia ohjelmistoja
            - Elan[98], Praat[99]
    - Annotointi- ja litterointisuositukset:
        - TEI?
        - Puheen litterointi
            - IPA, SU[100] ja CA[101]-systeemit
                - yhtenäinen ohjeistus mahdolliseen foneettiseen transkriptioon litteraatin ohella (tai sen asemasta)?
        - Sentimentti
            - 5 portainen asteikko?
            - Tekstissä/lauseessa voi olla läsnä useita sentimenttejä
            - Mihin sentimentti kohdistuu?
            - Sarkasmin tunnistaminen
- Jonkin annotaatioalustan tai alustojen edelleen kehittäminen tai uuden tekeminen
    - Kannattaako jotain olemassa olevia edelleen kehittää vai tehdä kokonaan uusi?
    - Kuka kehittää eteenpäin tai tekee uuden?
- Onko mahdollista luoda toimiva joukkoistamisympäristö?
    - Esim. mikromaksuja palkkiona suomenkielisen puheen litteroinnista
    - On huomioitava, että satunnaisten litteroijien tuottama litteraatio saattaa olla hyvinkin heikkolaatuista, jolloin samaan kieliainekseen pitää validoinnin vuoksi käyttää useita litteroijia

[98] https://www.kielipankki.fi/tuki/elan/
https://tla.mpi.nl/tools/tla-tools/elan/
[99] https://www.kielipankki.fi/tuki/praat/
[100] Onko liian subjektiivinen?
[101] Sopiiko kieliteknologisin tarkoituksiin lainkaan?

6.7. Kieliaineistoista lasketut mallit

Seitsemännessä operaatiossa lasketaan malleja jo olemassa olevilla välineillä jo saatavilla olevista aineistoista. Tarkoitus olisi varmistaa olemassa olevien ja valmiiksi laskettujen mallien käytettävyys (lisenssit ja saatavuus) yrityskäytössä ja laskea uusia malleja saatavilla olevista aineistoista. Aineistojen tekijänoikeudet tai tietosuojavaatimukset eivät yleensä välity niistä laskettuihin malleihin, joten kaikki lasketut mallit voidaan vapaasti jakaa CC0-lisenssillä.

Mallit kuvaavat aina niitä aineistoja, joista ne on laskettu. Riippuen käyttökohteesta tarkempi malli voi olla merkittävästikin parempi kuin iso laajasta aineistosta laskettu malli. Yksi projektin päätehtävistä onkin koota erilaisia aineistokokoelmia, joista malleja lasketaan. Tekstiaineistoista lasketuista malleista erityisesti BERT-mallille[102] on tällä hetkellä kysyntää. Mahdollisesti myös useammalla kielellä opetetut ”transfer”-oppimiseen[103] tarkoitetut mallit voisivat olla mahdollisia (multilingual BERT[104], XLM[105], XLNet[106]). Sekä FastText- että Word2Vec-malleja käytetään tällä hetkellä yleisesti, mutta niitäkään ei ole tarjolla kovin montaa erilaista. Puheaineistojen puolella kyseeseen tulevat lähinnä eri puheentunnistusohjelmistoille lasketut valmiit kielimallit (erityisesti Kaldi-alusta[107] on tällä hetkellä suosittu). Projektista tulisi kehittää prosessi, joka jatkaa tarvittavien mallien identifiointia tulevaisuudessa ja tuottaa malleja uusista aineistoista niiden kertyessä.

Työhön kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Mitä malleja tarvitaan?
    - Iso BERT malli kaikesta käytettävillä olevasta suomenkielisestä tekstistä
    - Pienempiä BERT malleja tarkemmin määritellyistä tekstiaineistoista
        - Tekstit eri aikakausilta (Kansalliskirjaston vanhoilla aineistoilla ja ilman)
        - Tekstin alkuperän perusteella jaotellut aineistot
            - Someaineisto, OCR:tty aineisto, virkateksti, lakiteksti, jne.
        - Eri genrejen perusteella jaetut aineistot

[102] Bidirectional Encoder Representations from Transformers: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
[103] http://ruder.io/state-of-transfer-learning-in-nlp/
[104] https://arxiv.org/pdf/1906.01502.pdf
[105] https://towardsdatascience.com/xlm-enhancing-bert-for-cross-lingual-language-model-5aeed9e6f14b
[106] https://mlexplained.com/2019/06/30/paper-dissected-xlnet-generalized-autoregressive-pretraining-for-language-understanding-explained/
[107] https://kaldi-asr.org

        - Myös yleiset FastText ja Word2Vec mallit kannattaa tehdä aineistoista samalla, koska aineistokokoelmien tarkempi määrittely ja kerääminen on tässä kuitenkin manuaalinen työ
    - Puheentunnistusmalleja eri aineistoista
        - Onko yleinen suomenkielinen puheentunnistusmalli mahdollinen/järkevä?
        - Avoimesti saatavilla oleva puheentunnistimen suomenkielinen kielimalli (yksi iso) tai kielimalleja (esimerkiksi eri murteille) käytettäväksi yhteensopivien puheentunnistimien kanssa.
- Kuka voi laskea?
    - Mallit hyödyllisiä myös tutkimuksessa ja niiden julkaiseminen CC0:na pitäisi olla ongelmatonta?
        - yhteistyössä yliopistojen ja CSC:n kanssa?
        - Tutkimuskäytössä on erittäin suuria datamääriä, joista malleja voidaan laskea

6.8. Avoimet ohjelmistokomponentit

Kahdeksannessa operaatiossa varmistetaan suomen kielen käsittelyyn tarvittavien ohjelmistokomponenttien saatavuus ja käytettävyys. Tarkoitus on tehdä avoimella lisenssillä (MIT tai CC0:aa vastaava) tuotantoversioita olemassa olevista kielenkäsittelyyn tarkoitetuista prototyypeistä tai menetelmäkuvauksista. Yhtenä päämääränä on olemassa olevien, tutkimuksen ohessa tehtyjen prototyyppien (ei teolliseen käyttöön soveltuvien) ja komponenttien tuotteistaminen ja niiden käytön mahdollistaminen eri ympäristöissä. Vaikka tutkijoiden kehittämät prototyypit olisivat käyttökelpoisiakin, tarvitaan niiden hankkimiseksi usein yhteydenottoja yksittäisiin tutkijoihin tai niitä saa käyttöönsä vain hankalien käyttölupaprosessien kautta. Palveluiden ja tuotteiden pilotointia varten olisi hyvä olla saatavissa nopeasti ja helposti käyttöönotettavia kielikomponentteja.

Monet kieliteknologiset ohjelmistokomponentit tai niiden prototyypit on kirjoitettu viime aikoina Python-kielellä, mutta Python-kieliset ohjelmat saattavat olla kymmeniä kertoja hitaampia kuin perinteisemmillä kielillä kirjoitetut ohjelmat.[108] Morfologiset analysaattorit ja syntaktiset jäsentimet suomelle ovat perinteisesti pitkälle kehitettyjä akateemisen yhteisön tuella, mutta myös niiden saatavuus (CC0 tai CC-BY) ja toteutuksen tuotantokelpoisuus käyttäen jotakin tehokasta ohjelmointikieltä (esimerkiksi C, C++ tai Java) olisi syytä varmistaa. Tällä hetkellä yrityksissä on käytössä esimerkiksi yleisesti vapaana lähdekoodina saatavissa oleva Voikko-ohjelmisto.[109]

[108] https://greenlab.di.uminho.pt/wp-content/uploads/2017/09/paperSLE.pdf
[109] https://voikko.puimula.org

Puheentunnistin, puhesynteesi sekä OCR -ohjelmistot ovat tällä hetkellä tuotteistettavien komponenttien listalla. Puheentunnistimessa tarvitaan lähinnä suomenkielisiä malleja, jotka tulevat kuudennesta projektista, mutta myös täysin käyttövalmis yleinen ohjelmistopaketti pitäisi koostaa. Myös NER-, sentimentti- ja tunnelatausanalysaattorit olisivat tällä hetkellä toivottuja. Ohjelmistokomponenttien käytettävyys tulisi varmistaa myös tarpeen mukaan uudelleenkirjoittamalla dokumentaatiota tai käyttöohjeita. Jo tehtyjen ohjelmistokomponenttien ylläpitoon, jatkuvaan jakeluun sekä uusien prototyyppien tuotteistamiseen tarvitaan prosessi tai organisaatio käyttäen mahdollisuuksien mukaan hyväksi ensimmäisen operaation tuloksia.

Työhön kuuluisivat esimerkiksi seuraavat selvitykset ja toimenpiteet:

- Tuotantokelpoisia ohjelmistokomponentteja
    - Ohjelmistokomponentit tehokkaina kirjastoina (esimerkiksi C, C++ tai Java) ja näitä kirjastoja käyttävien APIen tekeminen muihin kieliin kuten Pythoniin
- Dokumentointi ja käyttöohjeet
- Mitkä kannattaa tehdä ja missä järjestyksessä?[110]
- Kuka tekee?
- Miten näiden kehittymistä seurataan jatkossa?

7. Yhteenveto

Suomenkielisen tekoälyn kehittämisohjelman seuraavaa vaihetta varten olemme tässä esiselvityksessä keränneet aiheita operaatioihin, jotka eivät sisällä suuria määriä uutta tutkimusta tai uusien teknologioiden kehittämistä, mihin on tällä hetkellä jo olemassa rahoituskanavia. Nyt esitettyjen operaatioiden tarkoituksena on ennen kaikkea luoda mahdollisuudet huolehtia suomenkielisistä kieliresursseista niin, etteivät teknologiaa ja digipalveluita kehittävät yritykset ja yleishyödylliset toimijat joutuisi tekemään kompromisseja johtuen kieliresurssien puutteista. Tähän tarvitaan uuden toimijan perustaminen ja/tai olemassa olevan organisaation tukeminen, kuten on esitetty ensimmäisessä operaatioaihiossa (kappale 6.1).

Suosituksissa ehdotetut operaatiot muodostavat laajan kokonaisuuden niin, että niistä kaikki hyötyvät toisistaan ja kaikkia tarvitaan laajan kokonaishyödyn saavuttamiseksi.

[110] Ei kaikkia yhtäaikaisesti alussa, vaan valitaan ensin tärkein ja arvioidaan toteutuksen toimivuus ennen kuin siirrytään seuraaviin komponentteihin. Eli arvioidaan onko tuotantomalli toimiva.

Tämän esiselvityksen suosituksena on, että saavuttaakseen ja varmistaakseen täysimittaisesti potentiaalisen yhteiskunnallisen hyödyn, kaikki ehdotetut kahdeksan projektia olisi syytä toteuttaa pikimmiten, ja resurssien kehitykseen ja ylläpitoon tarvitaan ehdotettu neutraali organisaatio. Kunkin operaation kustannukset, arvioitu kesto, ja saatavat hyödyt on vielä syytä tarkemmin selvittää.

(Yksittäisiä) operaatioita ja niiden päämääriä voi ja tuleekin muokata projektien alettua ja uusien ja tarkempien näkökulmien esille tullessa. Toteutuakseen kunnolla, jokainen operaatio tarvitsee vahvan projektinhallinnan, jolla on vahva ymmärrys ja kokemusta liiketoiminnan kehittämisestä, projektityöskentelystä ja operaation kohteena olevasta kieliresurssista – kappaleessa 6.1 kuvatussa operaatiossa erityisesti liiketoiminnallinen ja johtamisosaaminen korostuvat.

Toteutuksessa ja sen hallinnassa tulee ymmärtää suomenkielisen tekoälyn kehittämisohjelman kokonaistavoitteet ja kyetä hyvin toimimaan yhdessä eri operaatioiden ja tahojen kanssa kokonaisuuden hyödyn varmistamiseksi.

Liite 1. Esiselvityksen työstämiseen osallistuneet tahot

Tämä lista sisältää kaikki tahot ja henkilöt, joille jokin versio esiselvityksestä on lähetetty kommentoitavaksi. Aivan kaikki listalla olleet eivät ennättäneet kommentoida esiselvitystä ennen sen valmistumista, eikä esiselvitys ei välttämättä millään tavoin kuvasta listalla olevien tahojen mielipiteitä esiselvityksessä käsiteltyihin asioihin. Osa listalla nimetyistä henkilöistä osallistui myös itse esiselvityksen kirjoitustyöhön, josta suuret kiitokset itse kullekin.

- Aalto yliopisto / Mikko Kurimo, Osmo Kuusi
- Aivan.ai / Jussi Karttila, Ville Laurikainen
- Alma Media / Santtu Elsinen
- Bitville Oy / Antti Keurulainen, Jouko Ranta
- Business Finland / Outi Keski-Äijö, Aki Parviainen
- CSC / Martin Matthiesen
- DAIN studios / Ulla Kruhse-Lehtonen
- DNA / Erkka Ryynänen, Kati Sulin, Pekko Ojanen
- DoubleVerify / Petrus Pennanen
- Eduskunnan kanslia / Sari Wilenius
- Etuma – Conexor / Pasi Tapanainen
- Feelingstream / Riikka Kokko, Risto Hinno, Terje Ennomäe
- Finnair / Patrik Etelävuori, Juha Karstunen, Minna Kärhä
- Fonecta – 020202 / Anders Gustavsson
- Fujitsu / Antti Suni, Jari Vuori
- Futurice / Claes Kaarni, Paavo Punkari
- Futusome / Sonja Baer
- GetJenny / Teemu Kinos
- Helsingin yliopisto, FIN-CLARIN, Kielipankki / Tommi Jauhiainen, Krister Lindén, Jyrki Niemi, Mietta Lennes, Hanna Westerlund
- Helsinki Intelligence Oy & TAIKA AI / Ville Henttonen
- IBM Finland / Maarit Palo, Niina Levo, Sara Elo Dean, Jukka Ruponen
- Iloom Oy / Sari Siikasalmi
- Inscripta / Simo Sorsakivi
- Kansalliskirjasto / Jussi-Pekka Hakkarainen, Osma Suominen, Nicholas Volk
- Kela / Ville Viitasaari, Janne Pulkkinen, Riitta Savolainen-Mäntyjärvi, Mika Saastamoinen, Heli Knihti, Vesa Kaakkuriniemi
- Kielikone / Arto Leinonen
- Kotus / Ulla-Maija Forsberg, Lotta Jalava

- LeadDesk / Jarno Tenni
- Lingsoft / Sebastian Andersson, Janne Vainikainen, Juha Tarvainen, Juhani Reiman
- Maahanmuuttovirasto / Jouko Salonen, Vesa Hagström
- Onerva hoivaviestintä Oy / Lauri Lehtovaara
- Osuuspankki / Hugo Gävert, Kristian Luoma
- Reaktor / Tiina Härkönen, Tommi Asiala, Nina Haukkovaara
- Sanoma / Vesa Lindqvist
- SFG Yhtiöt / Harri Fagerholm
- Silo.AI / Peter Sarlin, Filip Ginter
- Sofor / Matti Ruuskanen, Helena Malinen, Seppo Salo
- Sometrik / Olli Parviainen
- Speechgrinder / Otto Söderlund
- Teknologiateollisuus ry / Alexander Törnroth
- Tieto / Ari Rantanen
- Traficom / Kirsti Laurila
- Turun yliopisto / Tommi Kurki, Veronika Laippala, Osmo Kuusi
- UltimateAI / Reetu Karjalainen
- University of Alberta / Antti Arppe
- Utopia Analytics / Kari Kemppi, Mari-Sanna Paukkeri
- Vake / Terhi Marttila, Tuomas Teuri, Pia Erkinheimo
- Verohallinto / Mikko Laakso, Jarno Tuimala
- Voikko / Harri Pitkänen
- Wapice Oy / Markus Mäkelä, Mickey Shroff
- Yle / Aleksi Rossi, Anna-Leena Lappalainen, Ville Alijoki, Seija Aunila, Jukka-Pekka Heiskanen, Väinö Ala-Härkönen, Jarno M. Koponen
- Ääni Company / Janne Räsänen
- Yksityiset asiantuntijat: Teemu Ruokolainen, Viljami Venekoski